Research on Early Risk Predictive Model and Discriminative Feature Selection of Cancer Based on Real-world Routine Physical Examination Data

Guixia Kang
Key Laboratory of Universal Wireless Communications,
Ministry of Education
Beijing University of Posts and telecommunications
Beijing, China
gxkang@bupt.edu.cn

Zhuang Ni
Key Laboratory of Universal Wireless Communications,
Ministry of Education
Beijing University of Posts and telecommunications
Beijing, China
king_ni723@163.com

***Abstract*—most cancers at early stages show no obvious symptoms and curative treatment is not an option any more when cancer is diagnosed. Therefore, making accurate predictions for the risk of early cancer has become urgently necessary in the field of medicine. In this paper, our purpose is to fully utilize real-world routine physical examination data to analyze the most discriminative features of cancer based on ReliefF algorithm and generate early risk predictive model of cancer taking advantage of three machine learning (ML) algorithms. We use physical examination data with a return visit followed 1 month later derived from CiMing Health Checkup Center. The ReliefF algorithm selects the top 30 features written as Sub(30) based on weight value from our data collections consisting of 34 features and 2300 candidates. The 4-layer (2 hidden layers) deep neutral network (DNN) based on B-P algorithm, the support machine vector with the linear kernel and decision tree CART are proposed for predicting the risk of cancer by 5-fold cross validation. We implement these criteria such as predictive accuracy, AUC-ROC, sensitivity and specificity to identify the discriminative ability of three proposed method for cancer. The results show that compared with the other two methods, SVM obtains higher AUC and specificity of 0.926 and 95.27%, respectively. The superior predictive accuracy (86%) is achieved by DNN. Moreover, the fuzzy interval of threshold in DNN is proposed and the sensitivity, specificity and accuracy of DNN is 90.20%, 94.22% and 93.22%, respectively, using the revised threshold interval. The research indicates that the application of ML methods together with risk feature selection based on real-world routine physical examination data is meaningful and promising in the area of cancer prediction.**

Keywords—Early cancer diagnosis; ML algorithms; Discriminative feature selection; Real-world routine physical examination data; The fuzzy interval of threshold

I. Introduction

According to the World Health Organization (WHO) statistics in 2012, cancer, known as a major cause of death across the world, is considered to cause about 8.2 million deaths more than all coronary heart disease or all stroke did. In addition, it is predicted that the number of new probable cases with cancer will rise from 14 million by 70% over the next 2 decades [1]. Measures should be taken in response to the urgency of the rising incidence of cancer. Cancer is a complex, multifactorial, and deadly disease which is really hard to diagnose. Most cancers at early stages show no obvious signs and symptoms. Moreover, almost 80% of patients are diagnosed at middle or advanced stages when curative treatment is not an option any more. The mortality and disability rate in cancer patients which is a difficult problem to overcome is a major public health concern. Scientists made efforts and tried different ways to detect kinds of cancer in the early stage, which could improve the probability of survival and reduce healthcare costs. Making accurate predictions for the risk of early cancer is a really challenging but meaningful work for every physicians.

As new technologies in the domains of healthcare develop rapidly, medical examination data are saved by the digital format named electronic health records (EHRs) and large numbers of cancer data have been stored which are used for guidelines for medical research. EHRs [2-4] provide details about patients, such as demographic information, history, heath status, medical condition, treatments and so on, which is beneficial for building precise early cancer risk prediction models, making personalized medicine decision and optimizing therapies for a specific patient.

Considering the widely proliferation and innovation of Machine Learning (ML), the application of ML algorithms in diseases diagnosis and personalized medicine using EHR data is becoming the focus in medical research field and is attracting much scientist's attention. Based on [5], logistic regression [6] is the most common way in diseases diagnosis. Varieties of ML methods such as Random Forest (RF), Decision Tree (DT) and Support Vector Machine (SVM) have been widely applied in cancer predictive models. Eshlaghy et al. [7] used three machine learning techniques on Iranian Center for Breast Cancer (ICBC) dataset to predict breast cancer recurrence and made comparisons of the performance of DT, SVM and ANN through accuracy and specificity, sensitivity. Eventually, DT had the lowest accuracy. Miotto et al. [8] proposed a deep feature learning way to derive the patient representation from raw EHR data and applied RF classifiers to predict the probability that patients develop a certain disease from 78 diseases including 14 cancers. Several studies [9-11] discussed microarrays and gene

978-1-5090-1610-5/16/$31.00 ©2016 IEEE

expression signatures and listed the limitations of them for the cancer predictive model. There are several challenges regarding risk prediction models for cancer: (i) those cancer risk predictive models utilize medical trials databases which have smaller size of patient sample, lacking of heterogeneity and can't represent the health condition of real-world patients fully. (ii) These models include medications information and uncommon variables which is not easily accessible for those people who need the risk predictive models. It is essential to build risk predictive models for cancer based on EHR data derived from routine physical examination rather than clinical trials databases.

Routine physical examination includes demographic information, vitals, blood routine examination, urine routine examination, biochemical examination, alpha fetoprotein (AFP) and carcino-embryonic antigen (CEA). All variables in physical examination list are considered as potential risk features for predicting the risk of cancer. We will dig into the relationship between every features and cancer to find discriminative risk features and exclude uninformative factors. In this article, we mainly focus on the routine physical examination used for constructing an early cancer risk predictive model for query patients. A return visit was followed 1 month after the physical examination for all the candidates. A feature selection method called ReliefF algorithm is used to calculate the weight of each risk feature and then some feature subsets are constructed for classifiers. Some ML algorithms (DNN, SVM, DT) are chosen to generate an early cancer risk predictive models derived from routine physical examination data. Finally, the diagnostic feedback will be generated and sent to candidates by call, SMS or the Internet in accordance with predictive results from risk prediction model and candidate's individual information from Health Checkup Center. The whole architecture of our work is displayed in Fig.1.

Other parts of the paper are organized as follows: the study population in this paper we research on and the details of our dataset are presented in next part and Part 2 also makes a brief introduction of ReliefF algorithm and three chosen ML algorithms. In part 3, we have experiments of ReliefF method and classification algorithms which obtain the feature subsets and build risk predictive models, respectively, and make comparisons on the performance of three risk predictive model. We draw a conclusion of our entire work in part 4.

II. Matreials and Methods

A. Study Subjects

All routine physical examination data used by the early cancer risk predictive models we build are from databases of CiMing Health Checkup Center. We are authorized by CiMing Health Checkup Center to access the data for our research without any commercial purposes. There is still a challenge in front of us that is defining the cases and controls cohort of cancer. Here are the criteria we use to define the cases: (i) The candidate was diagnosed with cancer in 30-day interval after the routine physical examination; (ii) cancer diagnosis appeared on the corresponding list of ICD-9 codes; (iii) cancer appeared on at least two outpatient; (iiii) The age of candidate was 20~85 at the time of cancer diagnosis. The candidates were eligible as controls if they have at least one visit to a routine physical examination and have no cancer diagnosis in 30-day interval after the physical examination. Finally, a total of 650 cancer cases and 1650 cancer controls were selected from databases who met the criteria above.

B. Dataset Description

Routine physical examination data from CiMing Health Checkup Center databases includes demographic information, vitals, blood routine examination, urine routine examination, biochemical examination, alpha fetoprotein (AFP) and carcino-embryonic antigen (CEA). The detailed items in each categories were displayed:

- Demographic information including age, gender.

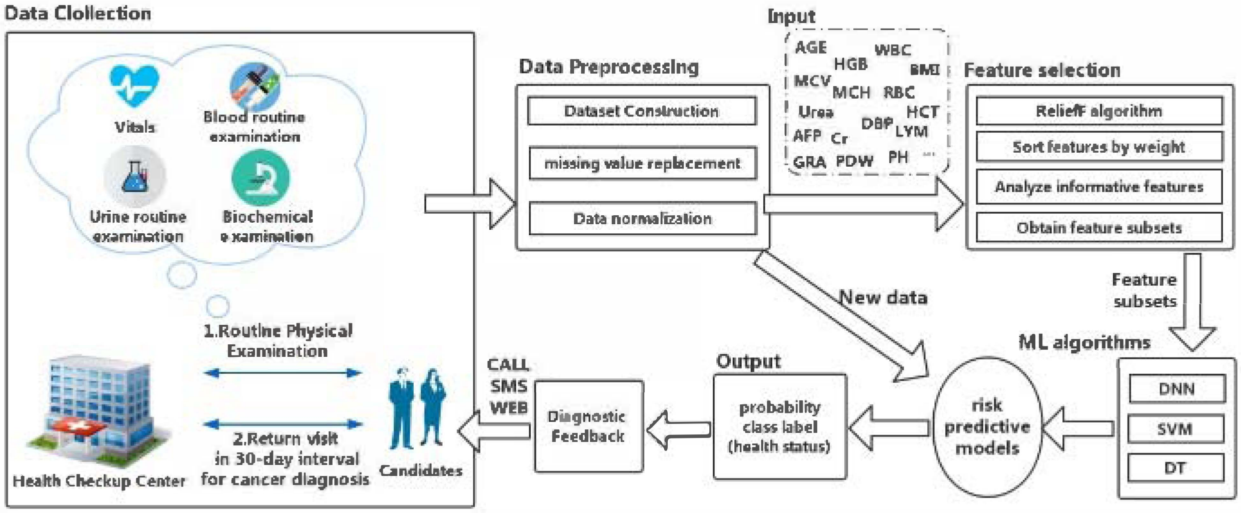

Fig.1.The whole framework of early risk predictive model and discriminative feature selection of cancer based on real-world routine physical examination data.

- Vitals including Body Mass Index (BMI), Systolic Blood Pressure (SBP) and Diastolic Blood Pressure (DBP) [5].
- Blood routine examination including White Blood Cell (WBC), Red Blood Cell (RBC), Hemoglobin (HGB), hematokrit (HCT), Mean Corpuscular Volume (MCV), Mean Corpuscular Hemoglobin (MCH), Mean Corpuscular Hemoglobin Concentration (MCHC), Red Blood Cell Volume Distribution Width Coefficient of Variation (RDW-CV), Red Blood Cell Volume Distribution Width of the Standard Deviation (RDW-SD), Platelet(PLT), Mean Platelet Volume (MPV), Platelet volume distribution width(PDW), granulocyte (%GRA),lymphocyte (%LYM) and Monocytes (%MON).
- Urine routine examination including Potential of Hydrogen (PH) and Specific Gravity (SG).
- Biochemical examination including Urea, Creatinine (Cr), Uric Acid (UA), Fasting Blood-Glucose (FBG), Alanine Aminotransferase (ALT), Triglyceride(TG) and Total Cholesterol(TC).
- Other risk predictors including alpha fetoprotein (AFP) and carcino-embryonic antigen (CEA).

The label (class variable) in our model is health status. If a candidate was diagnosed with cancer in 1-month interval after the physical examination, the label is 1, otherwise it is 0. There are a total of 34 risk predictor variables including 33 numeral variables which comprise age, vitals, blood routine examination, urine routine examination, biochemical examination, AFP and CEA and just 1 binary variable, gender.

Eventually, the dataset used for our model is composed of 34 features and 2300 candidates (650 cases and 1650 controls).

There are several missing values for some risk predictive variables in the dataset, which is a representative characteristic for real-world physical examination data. If the missing rate of the predictor is beyond 15%, the predictor will be excluded and otherwise the replacement of missing data should be considered. Multiple imputation (MI) is used to handle the missing values by the IBM SPSS Statistics 23 software. The normalization of variables is different by feature type. All the numerical variables in our dataset are normalized by Z-Score and the binary variables remain the same. Table I lists all the features of candidates in the dataset, the statistical parameters of the features, such as mean and standard and the number of missing data in every features. The formula of Z-Score is:

$$Z = (X-\mu) / \sigma \qquad (1)$$

TABLE I. DETAILS ABOUT FEATURES IN OUR DATASET

Categories	Features	No. of Missing Values	Mean ± Std（Units）
Demographics	Age(F1)	0	47.35±14.73 (year)
	Gender(F2)	0	1347 female (58.57%)
Vitals	BMI(F3)	42	24.27±3.78(Kg/m2)
	SBP(F4)	23	118.92±17.04(mmHg)
	DBP(F5)	23	76.76±10.75(mmHg)
Blood routine examination	WBC(F6)	4	5.93±1.49(10^9/L)
	RBC(F7)	4	4.68±0.49(10^{12}/L)
	HGB(F8)	4	140.80±15.67(g/L)
	HCT(F9)	4	41.56±6.54(%)
	MCV(F10)	4	90.26±4.80(fL)
	MCH(F11)	4	30.17±2.49(pg)
	MCHC(F12)	4	333.97±16.99(g/L)
	RDW-CV(F13)	10	12.63±1.30(%)
	RDW-SD(F14)	4	43.85±2.92(fL)
	PLT(F15)	4	224.78±58.58(10^9/L)
	MPV(F16)	4	9.77±1.03(fL)
	PDW(F17)	4	12.29±2.09(%)
	GRA(F18)	14	3.50±1.17(10^9/L)
	%GRA(F19)	4	58.41±7.86(%)
	LYM(F20)	4	2.04±0.58(109/L)
	%LYM(F21)	62	34.93±7.83(%)
	MON(F22)	66	0.39±0.16(10^9/L)
	%MON(F23)	67	6.66±2.04(%)
Urine routine examination	PH(F24)	42	5.95±0.70
	SG(F25)	35	1.021±0.006
Biochemical examination	Urea(F26)	58	5.07±1.37(mmol/L)
	Cr(F27)	43	64.03±15.60(μmol/L)
	UA(F28)	16	300.44±86.56(μmol/L)

	FBG(F29)	5	5.31±1.14(mmol/L)
	ALT(F30)	4	24.58±23.13(U/L)
	TG(F31)	3	1.56±1.50(mmol/L)
	TC(F32)	3	5.01±0.98(mmol/L)
Other risk predictors	AFP(F33)	0	6.07±34.60(μg/L)
	CEA(F34)	0	4.95±41.26(ng/mL)

In the formula above, X is the raw value of numerical variables. Specifically, μ and σ is the mean and standard deviation before normalization. Z, the value after normalization, is positive when the raw value is bigger than the mean.

C. Feature Selection Algorithm

In the field of medicine, a feature subset with less variables means less physical examination items and lower medical expense. However, A great number of features are collected before building the predictive model but not all the variables are informative and useful [12]. It is imperative to eliminate the redundancy of the features and select more informative variables for increasing the accuracy and efficiency of the predictive model. In our model, a feature selection algorithm called ReliefF [13] was applied to our risk predictive variables. Before implementing the ReliefF algorithm, the missing data should be filled up by the MI. ReliefF aims to calculate the weight of every features based on whether they can differentiate between close instances better than others. If the weight of the feature is bigger, the feature will be more discriminative, have stronger predictive power and make more contributions to the accuracy of the model. Comparing with wrapper methods, we prefer to get each feature's relation to label. The process of ReliefF is presented:

Input: Training dataset D (every training sample in D is a vector of both features and label), m is the number of sampling defined by user, k is the number of neighbor.

Step1: Randomly select an instance Hi (i = 1, 2, …, m) from dataset D, find k nearest neighbors respectively from the same class, denoted as Sj. The distance between instances we use is Euclidean Distance. From class C (not equal to class(Hi)), find k nearest neighbors from the opposite class, denoted as Oj(C).

Step2: Update the weight vector W[F] for all features The formula of W[F] is defined as (2).

$$\mathrm{W[F]} := \mathrm{W[F]} - \sum_{j=1}^{k} diff(F,Hi,Sj)/(\mathrm{m} * \mathrm{k}) + \sum_{C \neq class(Hi)}\left[\frac{P(C)}{1-p(class(Hi))}\sum_{j=1}^{k} diff(F,Hi,Oj(C))\right]/(m * k) \quad (2)$$

Output: the weight vector W[F].

The function diff(F,Hi,Hj) is used to operate the subtraction between the values of the feature F for two samples Hi and Hj. Function diff(F,Hi,Hj) is calculated as:

$$\mathrm{diff}(\mathrm{F},\mathrm{H_i},\mathrm{H_j}) = \begin{cases} 0, & F \text{ is binary and } \mathrm{Hi[F]} = Hj[F] \\ 1, & F \text{ is binary and } \mathrm{Hi[F]} \neq Hj[F] \\ \frac{|Hi[F]-Hj[F]|}{\max(\mathrm{F})-\min(\mathrm{F})}, & F \text{ is numerical} \end{cases} \quad (3)$$

D. ML Algorithms

To predict whether candidates are diagnosed with cancer in 30-day interval after routine physical examination, we apply three data mining techniques, namely DNN, SVM, DT to generate early risk predictive model. Next, we will make a brief introduction to these three data mining techniques.

[1] DNN: Deep neutral network [14,15,16] is an efficient deep learning algorithm developed to simulate the function and structure of human being brain which is comprised of neurons interconnected by axons. The interconnected neutron is a basic operating element of the deep neutral network and the strength of interconnection between neutrons is represented by the weights matrix updated by Back Propagation (BP) algorithm. The neutron takes the inputs X (n-dimensional vector) and a bias element bj as its input and then calculate the multiplication of the inputs X for the weights matrix W added by the bias bj to get aj called the activation state and eventually outputs the binary value yi after implementation of activation function f(•) with a user-defined threshold value θ. The formula of activation state aj is displayed as followed:

$$\mathrm{a_j} = \sum_{i=1}^{n} x_i \mathrm{w_{ij}} + \mathrm{b_j} \quad (4)$$

There are several popular activation functions which transfers the output value of every single neuron to the value between [0, 1] or [-1, 1] such as sigmoid function, tanh function and so on. In our model, we choose sigmoid function, which is a non-linear function and is the most common activation function in neutral network. The formula is presented below:

$$\mathrm{f(x) = 1/(1+exp(-x))} \quad (5)$$

The curve of the sigmoid is like "S" and the range of output value is between 0 and 1. The consequence of the output y_i is (6):

$$\mathrm{y_i} = \begin{cases} 1, & f(a_i) \geq \theta \\ 0, & f(a_i) < \theta \end{cases} \quad (6)$$

In our paper, θ is set as 0.5. Both the number of neutron in every layer and the number of hidden layer are combined to affect the efficiency and the prediction accuracy of deep neutral network. The two parameters are optimized by grid search method which is verified by 5 fold Cross-Validation (CV) in part 3.

[2] Support Vector machine: SVM [16,17,18] is more and more popular as a method of ML techniques applied in the area

of disease prediction in terms of its ability of processing non-linear data and precise predictive performance. SVM maps the original input vectors into a higher feature dimensional space using different kernel functions, which makes input data linear separable possibly and then identifies the optimal hyperplane considered as decision boundary with the maximum margin that divides the data points into two clusters. Fig.2 displays the graph of SVM in hyperspace. In our experiments, LIBSVM, an open toolbox for SVM designed by Chih-Jen Lin from National Taiwan University [19], is used in the MATLAB R2014b. We compare the performance of different kernel functions and eventually choose SVM with the linear kernel function to train the model verified by 5-fold CV, which accomplishes high generalizability and thus is reliable for prediction of new dataset.

[3] Decision Tree: CART(Classification and Regression Trees) is a widely rule-based classification approach that builds a binary tree through iteratively partitioning the subset (called a leaf) into two subsets (called sub-leaves) according to the minimization of criterion calculated on the resulting sub-leaves until the entire tree is completed. Each split is in accordance with a variable and not all variables may be used. Each sub-leaf is then split based on decision rules[20]. We do experiments of several different DT algorithms such as CHAID, QUEST and CART in IBM SPSS Statistics 23 to analyze the risk and finally choose the CART to construct the risk predictive model.

III. Experiments and Discussions

In this part, the experiments of the early risk predictive model of cancer is presented, which include 5 main respects: (1) analyzing discriminative features and generating several different feature subsets for ML techniques according to the weight value of every features which is calculated by ReliefF algorithm, and the feature will be excluded if its weight value is below the threshold value defined by user. (2) Finding the optimal parameters (the number of neutron in every layer and the number of hidden layer) for DNN. (3) Constructing several early risk predictive model of cancer with different feature subsets based on 3 ML algorithms, and evaluating the performance of these different classifier models in comparisons. (4) 5-fold cross validation applied to all three models is a good method to alleviate the overfitting problem and enhance the generalization ability of the model. (5) Proposing the fuzzy interval of threshold in DNN to improve the predictive accuracy of model.

Some metrics are proposed to evaluate and measure the performance of the feature subsets and early risk predictive model of cancer. For the evaluation and effectiveness of the feature subsets, predictive accuracy is in the first place and then dimensionality reduction is considered. For assessing the performance of risk predictive model based on three ML algorithms, Receiver operating characteristics (ROC) curves for DNN, SVM, DT models are plotted and the average area under curve (AUC) value that is sum of ROC curve of 5-fold cross validation of models is computed. Statistical parameters (accuracy, sensitivity and specificity) are another metrics to assess the performance of models.

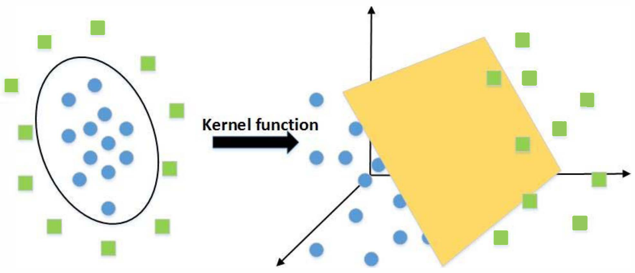

Fig.2. The graph of SVM in hyperspace.

A. Experiments for discriminative feature selection

The consequences of discriminative feature selection according to ReliefF method are presented below. Fig.3 displays the weight value of all features calculated by ReliefF algorithm. The order of features weight from high to low which can be easily seen from Fig.3 are F22, F9, F21, F18, F23, F19, F6, F1, F8, F7, F25, F20, F4, F17, F10, F27, F29, F28, F14, F13, F24, F11, F5, F16, F15, F32, F12, F26, F3, F34, F2, F30, F33, F31. We set up the threshold (=0.2) to exclude some uninformative feature (F2, F30, F33, F31). In our experiments, these 4 features (gender, ALT, AFP, TG) are less discriminative and provide little information for cancer prognosis. The top 30 features obtained by ReliefF are considered as informative features.

However, the exact number of features needed to achieve the best performance is still unknown. Therefore, these 30 informative features are selected to generate feature subsets. 30 subsets with different attributes are constructed to build DNN architectures. Sub(n) means that this feature subset includes the top n of features based on weight. Table II shows part of 30 subsets with discriminative features based on the weight value.

TABLE II. The Part Of 30 Subsets With Discriminative Features Based On The Weight Value

Subsets	**Features**	**No. of informative Features**
Sub(1)	F22	1
Sub(2)	F22,F9	2
Sub(3)	F22,F9,F21	3
…	…	…
Sub(9)	F22,F9,F21,F18,F23,F19,F6,F1,F8	9
Sub(10)	F22,F9,F21,F18,F23,F19,F6,F1,F8,F7	10
Sub(11)	F22,F9,F21,F18,F23,F19,F6,F1,F8,F7,F25	11
Sub(12)	F22,F9,F21,F18,F23,F19,F6,F1,F8,F7,F25,F20	12
…	…	…
Sub(29)	F22,F9,F21,F18,F23,F19,F6,F1,F8,F7,F25,F20,F4,F17,F10,F27,F29,F28,F14,F13,F24,F11,F5,F16,F15,F32,F12,F26,F3	29
Sub(30)	F22,F9,F21,F18,F23,F19,F6,F1,F8,F7,F25,F20,F4,F17,F10,F27,F29,F28,F14,F13,F24,F11,F5,F16,F15,F32,F12,F26,F3,F34	30

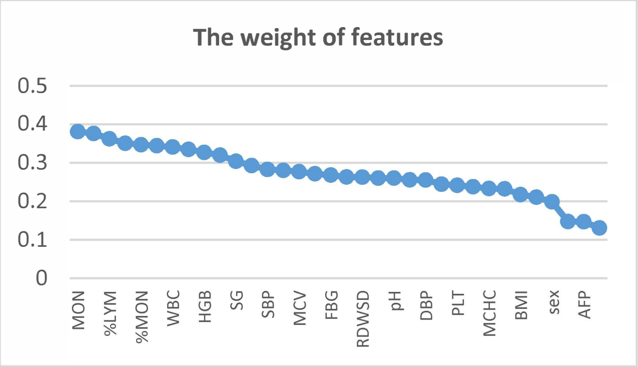

Fig.3. The weight of features sorted from high to low.

B. *Experiments for risk pridictive models*

In order to evaluate how accurately the risk predictive model presents at predicting if a candidate developed cancer, in this paper, we estimate the capability of the risk predictive model based on three ML algorithms to measure whether test candidates are possible to be diagnosed with cancer within 30-day interval. The entire dataset (650 cancer cases and 1650 controls) is separated into 5 portions which is different in content but is equal in size. In every iteration, 1 portion is chosen as the test collection in turn and the rest of these portions are used as training set to train the risk predictive model. The final results is the average value of 5 iterations. The effects on the prediction accuracy of the numbers of layers of DNN are described by the parameters (AUC and accuracy) in Fig.4. DNN with 2 hidden layers has the best performance (86.00% for accuracy and 0.882 for AUC) and no further progress is made for all metrics whose results is fairly stable after 2 hidden layers. In this case, we choose a deep neutral network including 2 hidden layers. We conduct experiments on these 30 subsets using DNN of 2 hidden layers in order to evaluate the predictive accuracy of feature subsets. Classification accuracy of models based on DNN of 2 hidden layers for different feature subsets is computed in Fig.5. The highest accuracy of models, 0.86, is based on Sub(30). To verify the correctness and effectiveness of ReliefF algorithm, we still build models using Sub(31), Sub(32), Sub(33), Sub(34). As a result, the accuracy of DNN Classifiers did not improve when these features (F2, F30, F33, F31) are added to Sub(30).

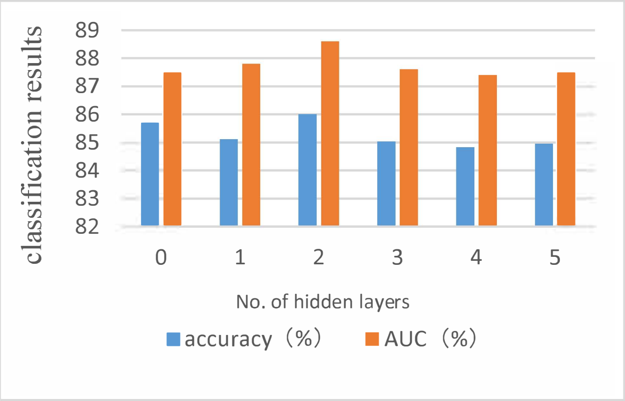

Fig.4 Evaluation of DNN based on the number of hidden layers.

Finally, we build a 4-layer (2 hidden layers) deep neutral network based on B-P algorithm as early risk predictive model of cancer in the platform of MATLAB R2014b. The number of neutron in the input layer is 30 which depends on the feature subsets we choose (here, we choose Sub(30)) and there is just a single node in output layer which outputs the health status of candidates. The number of neutron in first and second hidden layer (h1, h2) is optimized by grid search with 5-fold cross validation. Consequently, h1 equals to 43 and h2 is 10. The structure of deep neural network is exhibited by Fig.6 which is comprised of input layer, 2-hidden layer and output layer.

The dataset used as the input of three risk predictive models includes 30 features and 2300 examples (650 cancer cases), namely Sub(30).

The performance of 3 risk predictive models in respect of AUC-ROC and statistic parameters (accuracy, sensitivity and specificity) is compared. Fig.7 illustrate AUC-ROC of the three risk prediction models generated by DNN, SVM, DT classification algorithms. As is shown in Fig.7, the largest AUC value (0.928) is observed for SVM and the AUC for DT (0.824) is much lower. The results of three classifiers for early cancer risk prediction in 30-day interval are shown in Table III.

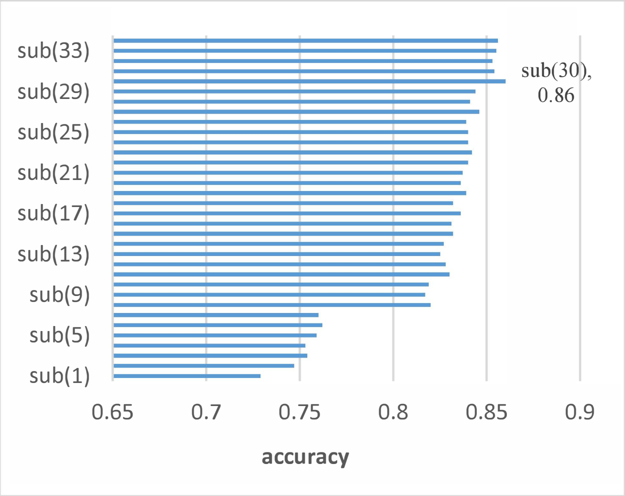

Fig.5. The performance of DNN models for 34 features subsets

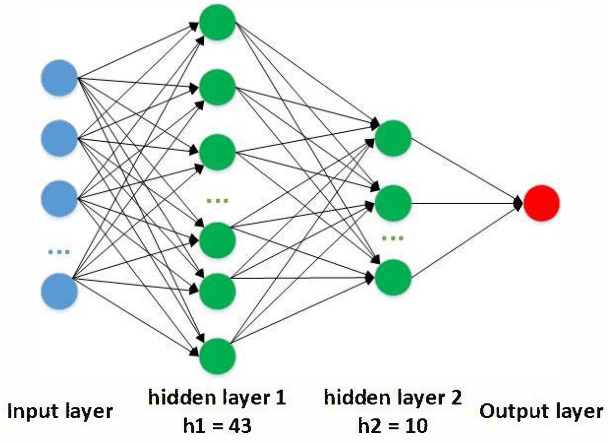

Fig.6. The structure of deep neutral network.

TABLE III. THE RESULT OF 3 CLASSIFIERS FOR EARLY CANCER RISK PREDICTION IN 30-DAY INTERVAL IN TERMS OF ACCURACY, SENSITIVITY AND SPECIFICITY

Algorithm	**Accuracy**	**Sensitivity**	**Specificity**
DNN	86.00%	64.07%	94.77%
SVM	83.83%	54.46%	95.27%
DT(CART)	83.60%	60.00%	91.50%

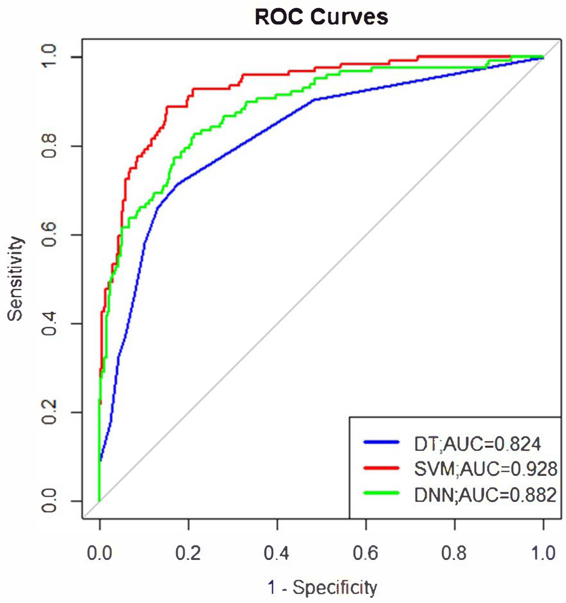

Fig.7. AUC-ROC of 3 risk prediction models generated by DNN, SVM, DT.

The performance criterion (accuracy, sensitivity) of DNN are better than those achieved by SVM and DT. Especially, DNN obtains accuracy of 0.86 with 3% improvement compared to DT method. The sensitivity of DNN is 9.61% larger than that of SVM. The specificity of SVM, 95.27%, which implies that it classifies almost all true negatives correctly is superior to DNN and DT. The consequences of these experiments illuminate that DNN is more suitable in generating early cancer risk prediction model taking accuracy and sensitivity into consideration but if the high AUC and specificity are preferred, SVM will be a better choice.

C. Experiments for the fuzzy interval of threshold in DNN

The threshold of DNN θ makes effects on predictive accuracy of model, so in the section we do some experiments of adjusting the threshold of DNN θ. Finally, We choose θ_1 equals 0.72 which means that when $f(a_i)$ is larger than 0.72, $y(a_i)$ is set as 1. And then we set θ_2 as 0.23 which means that $y(a_i)$ is 0, when $f(a_i)$ is lower than 0.23. The range between 0.23 and 0.72 is called "fuzzy interval" in which samples can not be classified . In this method, the performance has improved a lot and the sensitivity, specificity and accuracy of DNN is 90.20%, 94.22% and 93.22%, respectively. However, the limitation of this method is that some of samples will be located in the fuzzy interval where samples can not be anticipated. Only 61.5% of samples can be judged with the chosen thresholds in our dataset.

IV. CONCLUSIONS

In our research, we develop the 4-layer (2 hidden layers) deep neutral network based on B-P algorithm, the support machine vector with the linear kernel and decision tree CART for early risk prediction of cancer in 30-day interval after the routine physical examination. Real-world routine physical examination data acquired from databases of CiMing Health Checkup Center is used for training the models and constructing early risk prediction model of cancer. ReliefF algorithm is applied to calculate the weight of features and then obtain the significant and discriminative features which are more valuable and effective for cancer prediction. The proposed algorithms achieve the best performance based on Sub(30) which contains top 30 discriminative features. Some criteria including Accuracy, AUC-ROC, Sensitivity and Specificity are used for comparing the performance of 3 methods we built. The results reveal that different approach has its own superiorities and deficiencies. Taking account of accuracy and sensitivity, DNN is the optimal approach for early cancer risk prediction model but when it comes to AUC and specificity, SVM will be an option. Using the fuzzy interval of threshold in DNN can improve the performance of the predictive model derived from DNN but make some samples unclassified. Moreover, according to our results above, it is visible that the application of feature selection and ML classification methods based on real-world routine physical examination data will brings bright prospect in the area of cancer prediction.

ACKNOWLEDGMENT

This work was supported by National Science and Technology Major Project of China (No.2013ZX03006001), the National High Technology Research and Development Program of China ("863" Program, No.2015AA01A709) and National Natural Science Foundation of China (61471064).

REFERENCES

[1] World Health Organization. Global health observatory data repository. 2011. Number of deaths (World) by cause. Available from: http://apps.who.int/gho/data/node.main.CODWORLD?lang=en.Last accessed 31 October 2013.

[2] Castaneda C, Nalley K, Mannion C, Bhattacharyya P, Blake P, Pecora A, Goy A, Suh KS. Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. J Clin Bi, 2014, 5(1):1-16

[3] Xu H, Fu Z, Shah A, et al. Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases. AMIA Annu Symp Proc 2011, 2011:1564-72.

[4] Sun J, Hu J, Luo D, et al. Combining knowledge and data driven insights for identifying risk factors using electronic health records. AMIA Annu Symp Proc 2012, 2012:901–10.

[5] V Taslimitehrani, G Dong, NL Pereira, M Panahiazar, J Pathak. Developing EHR-driven heart failure risk prediction models using CPXR (Log) with the probabilistic loss function. J Biomed Inform, 2016; 60:260-269.

[6] S.C. Bagley, W. Halbert, B.A. Golomb, Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain, J Clin Epidemiol, 2001, 54(10):979–985.

[7] Eshlaghy AT, Poorebrahimi A, Ebrahimi M, Razavi AR, Ahmad LG. Using three machine learning techniques for predicting breast cancer recurrence. J Health Med Inform 2013, 4:124.

[8] Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records, 2016, Scientific Reports.

[9] Koscielny S. Why most gene expression signatures of tumors have not been useful in the clinic. Science Translational Medicine, 2010, 2(14):305-312.

[10] Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet, 2005, 365(9458):488–92.

[11] Van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, et al. A gene expression signature as a predictor of survival in breast cancer. New Engl J Med, 2002, 347(25): 1999-2009.

[12] A. Gheyas Iffat, S. Smith Leslie. Feature subset selection in large dimensionality domains, Pattern Recognition, 2010, 43(1):5–13.

[13] M. Robnik-Sikonja and I. Kononenko. Theoretical and empirical analysis of Relief and ReliefF. Machine Learning, 2003, 53(1-2):23–69.

[14] Ayer T, Alagoz O, Chhatwal J, Shavlik JW, Kahn CE Jr, Burnside ES. Breast cancer risk estimation with artificial neural networks revisited: discrimination and calibration. Cancer, 2010, 116(14): 3310-3321.

[15] Zangooei, M. H., Habibi, J., & Alizadehsani, R. Disease diagnosis with a hybrid method SVR using NSGA-II. Neurocomputing, 2014, 136(8):14–29.

[16] W. Dai et al., Prediction of hospitalization due to heart diseases by supervised learning methods. Int J Med Inform, 2014, 84 (3):189–197.

[17] Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol Jl. 2015, 13:8–17.

[18] Hui, L., Xiaoyi, L., Ramanathan, M. & Aidong, Z. Prediction and informative risk factor selection of bone diseases. IEEE-ACM T Comput Bi, 2015, 12(1):554–559.

[19] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011.

[20] L Bel, D Allard, JM Laurent, R Cheddadi, A Bar-Hen. CART algorithm for spatial data: Application to environmental and ecological data. Computational Statistics & Data Analysis, 2009, 53(8):3082-3093.

[21] Menéndeza L.A., de Cos Juez F.J., Lasheras F.S., Riesgo J.A.A. (2010) - Artificial neural networks applied to cancer detection in a breast screening programme. Mathematical and Computer Modelling, 2010, 52(7-8): 983-991.

[22] EJ Satya, N Chandrakar. Artificial neural networks as classification and diagnostic tools for lymph node-negative breast cancers. Korean J of Chem Eng, 2016, 33(4):1318-1324.

[23] Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assn, 2014; 21:221-30.