IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 18, 2019, accepted November 29, 2019, date of publication December 3, 2019,
date of current version December 17, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2957367

A New Hybrid XGBSVM Model: Application for Hypertensive Heart Disease

WENBING CHANG[1], YINGLAI LIU[1], XUEYI WU[2,3,4], YIYONG XIAO[1], SHENGHAN ZHOU[1], AND WEN CAO[5]

[1]School of Reliability and System Engineering, Beihang University, Beijing 100191, China
[2]Hypertension Center, Fuwai Hospital, Beijing 100037, China
[3]National Center for Cardiovascular Diseases, Chinese Academy of Medical Sciences, Beijing 100050, China
[4]Peking Union Medical College, Beijing 100730, China
[5]Department of Operational Performance Improvement, Peninsula Regional Medical Center, Salisbury, MD 21801, USA

Corresponding author: Shenghan Zhou (zhoush@buaa.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 71971013 and Grant 71871003, and in part by the Fundamental Research Funds for the Central Universities under Grant YWF-19-BJ-J-330.

ABSTRACT The changes in people's life rhythm and improvement in material levels that happened in recent years increased the number of people suffering from high blood pressure in the world. Therefore, as a cardiac complication of hypertension, the prevalence of hypertensive heart disease has increased annually, it has seriously endangered the safety of human life, and the effective prediction of hypertensive heart disease has become a worldwide problem. This paper uses the newly proposed XGBSVM hybrid model to predict whether hypertensive patients will develop hypertensive heart disease within three years. The final experiment proves that through this model, hypertensive patients can learn their risk of hypertensive heart disease within 3 years and then undergo targeted preventive treatment, thereby reducing the psychological, physiological and economic burden. This paper confirms that the machine learning can be successfully applied in the biomedical field, with strong real-world significance and research value.

INDEX TERMS XGBSVM, hypertensive heart disease, SVM, XGBoost, biomedicine.

I. INTRODUCTION

Over the past decade, cardiovascular disease has become the main cause of death in the world [1]. The World Health Organization report shows that in 2008, approximately 17 million people died of cardiovascular diseases, accounting for 30% of global deaths. Cardiovascular disease has the characteristics of high morbidity, high disability and high mortality, and even with the most advanced and perfect treatment methods, more than 50% of survivors of a cerebrovascular accident can't fully care for themselves. The prevalence of cardiovascular diseases has become a global problem. According to a survey by the American Heart Organization, the number of deaths due to cardiovascular disease in the United States accounted for 32.8% of the total number of deaths of the country in 2008. In other words, in the United States, one out of every three deaths is due to cardiovascular disease, with a daily death toll of 2200; on average, about every 40 seconds, a person loses his life due to cardiovascular disease [2]. In 2010, according to the National Health Interview Survey, the National Center for Health Statistics announced that 11.7% of people over 18 had heart disease, and 23.6% had hypertension. According to Turkey's National Disease Burden Study, cardiovascular and cerebrovascular diseases account for 36% of all deaths in Turkey [3]. Excluding developed countries, 80% of cardiovascular disease-related deaths occur in middle-income and low-income countries worldwide [4].

Blood pressure is one of the most important factors affecting cardiovascular disease. According to previous studies, 13% of cardiovascular deaths are caused by elevated blood pressure or hypertension [5]. With the gradual improvement of lifestyle, unhealthy eating and living habits are the primary causes of various diseases, the number of hypertensive patients is also rising rapidly in clinics. The transition from hypertension to hypertensive heart disease is a relatively slow process, but long-term maintenance of high blood pressure in hypertensive patients can easily lead to changes in heart structure and function, leading to the occurrence of hypertensive heart disease. In the world, the prevalence of hypertension in adults is 26.4%, including 40% in Spain, 29.6% in

The associate editor coordinating the review of this manuscript and approving it for publication was Alberto Cano.

Britain, 38.3% in Japan and 27.4% in Egypt [6]. According to epidemiological data in recent years, more than 200 million adults in China suffer from hypertension [7]. The prevalence of hypertension was 18.8% among residents over 18 years old, and 24% among middle-aged people in China [8]. Especially in rural and economically regressive remote areas, because of poor awareness and control of hypertension, the prevalence rate is higher [9].

At present, many researchers are studying cardiovascular diseases from a pathological perspective. Used R. predicted cardiovascular disease by lipid-related indicators [10], and Kunutsor S. predicted cardiovascular disease by circulating glutamine transferase [11]. Nagata M. predicted the mortality of cardiovascular diseases by proteinuria and renal function decline [12]. Skretteberg P. predicted the incidence of coronary heart disease by high density lipoprotein cholesterol [13].

The concept of entropy was first introduced by statistical thermodynamics. Information entropy was used to measure the uncertainty of random variables. In recent years, the application of entropy has been expanding, such as electronic measurement error, clinical quantitative diagnosis and analysis, optimization problems, design risk management and other fields [14]–[16].

With the emergence of big data, machine learning (ML) has been pushed to the forefront and has had a large impact, which has many applications in the biomedical field. Dolatabadi A. extracted time and frequency domain features from the electrocardiogram and used support vector machines (SVM) to predict coronary artery disease [17]. Sinha P. used SVM to create a new decision support system for predicting chronic kidney disease [18].Kaur K. proposed a method based on principal component analysis and SVM classification to predict heart disease, and proved that the method is suitable for disease prediction [19]. Ren Y. proposed a hybrid neural network that combines two-way long-term memory and self-encoding network. Using the data of 35,332 hypertensive patients, the model was used to predict renal disease in hypertensive patients and the final prediction accuracy was 89.7% [20]. Yu J. proposed a new collaborative filtering model CFNBC based on the naive Bayesian classifier to predict the correlation of potential RNA long non-coding diseases [21]. Vijayarani S. used naive Bayesian method to predict liver disease [22]. Shinde R. combines naive Bayesian method and k-means clustering algorithm to establish a heart disease prediction system [23].

Deep learning is now gradually applied to medical image analysis [24]. Decencière E. used machine learning to extract information from eyeball images and diagnose diabetic retinopathy [25]. Kermany D. uses a deep learning framework to create a diagnostic tool for screening patients with treatable blinding retinopathy [26].

Chen. T proposed the XGBoost algorithm in 2016 [27]; it has received wide attention. Zheng H. used the XGBoost method to obtain the characteristic importance of a power system [28]. Torlay L. used the XGBoost algorithm to analyze the neurophysiological characteristics of the five linguistic regions in the two hemispheres of the human brain, and finally, patients with epilepsy were classified. It is proven by experiments that the XGBoost algorithm not only has excellent performance but also obtains the importance of features [29]. Chen, W. proposed a weighted XGBoost algorithm to classify complex radar signals. The results of the experiments show that the algorithm outperforms other ML algorithms in the performance of test sets [30]. There are also some studies that combine XGBoost with other methods to solve problems [31], [32].

Support vector machines have many good attributes, such as reduced "dimensionality disaster" and better stability. Shalev S. solved the SVM optimization problem by using a random subgradient descent method. In large-scale text classification problems, the running efficiency is much higher than the traditional SVM algorithm [33]. Pasolli E. proposed a new active learning method for SVM classification, and the experimental results show the effectiveness of the algorithm [34]. Lapin M. studied SVM+ and Weighted SVM; this study shows that weight learning is effective, and explains the limitations of SVM+ [35]. Yin Y. combined SVM and PCA methods for visual tracking, which proves that the algorithm have high stability [36].

Hospitals produce a large number of new and potentially valuable data every day, but the previous scientific theory is not mature enough and has not been widely developed in the medical field, so these potentially valuable data have not been fully utilized. For heart disease, many studies only focused on one or two indicators related to it from the pathological perspective, and most of the studies were broad and did not specify the specific types of cardiovascular diseases. Therefore, this part of the study is imperfect, there is much follow-up work that needs to be conducted.

As mentioned above, the incidence of hypertension is very high, and there is a great possibility that it may cause hypertensive heart disease. For patients, there are lives threatened, and they also suffer from high treatment costs and tremendous mental stress. According to current research, it is meaningful to predict whether hypertensive patients will suffer from hypertensive heart disease in the next few years and to predict heart disease by combining entropy and artificial intelligence. Therefore, based on the experimental background of hypertensive heart disease, the XGBSVM method is used to predict whether hypertensive patients developed hypertensive heart disease within three years, and the validity, accuracy and stability of the XGBSVM method are also verified. Doctors can conduct targeted preventive treatment for hypertensive patients who are prone to hypertensive heart disease. This prediction will greatly reduce the economic and mental burden of patients and avoid the risk of major diseases.

The innovation of this paper is that, first, a hybrid XGBSVM model is proposed that eliminates the process of complex feature selection in traditional models, it use hypertensive heart disease as an experimental scenario to predict whether hypertensive heart disease will occur in hypertensive patients within three years. Second, a new index, improved normalized entropy, based on entropy information is proposed to evaluate the model. Finally, this paper applies entropy and machine learning in the field of biomedicine.

This paper is divided into four chapters. The first chapter introduces the research background, necessity, significance and innovation. The second chapter introduces the existing methods, new methods, model evaluation indicators, and the technical route of this study. The third chapter is about results analysis and discussion. The fourth chapter summarizes the whole paper and evaluates the importance of this study.

II. THEORY AND METHODOLOGY

A. RF,GBDT AND XGBOOST

Before introducing the three algorithms mentioned above, we first briefly describe the bagging algorithm, which is an integration technology for training classifiers on raw data sets by reselecting k new data sets with playback sampling. This algorithm uses a set of trained classifiers to classify the new samples and then counts the classification results of all classifiers by majority voting or averaging the output; the highest result category is the final label. This kind of algorithm can effectively reduce bias and variance.

Random forest (RF) is a kind of bagging algorithm [37], [38]. First, RF uses a CART decision tree as a weak learner, and the CART decision tree is based on the Gini coefficient to select features. When generating each tree, the selected features are randomly selected. Therefore, the randomness of the features is guaranteed. Because of the randomness, it is very useful to reduce the variance of the model, so the random forest generally does not need additional pruning, and it can obtain better generalization and anti-overfitting ability.

To summarize, the advantages of RF are as follows:

(1) Because of the ensemble algorithm, the accuracy of RF is better than that of most single algorithms.

(2) The test set performs well in solving some problems, and because of the introduction of randomness, the random forest does not easily fall into overfitting.

(3) RF has a certain anti-noise ability, for example, it is not sensitive to default values, so it has certain advantages over other algorithms.

(4) Because of the combination of trees, random forests can process nonlinear data and are nonlinear classification (fitting) models.

(5) RF can process data of high dimension without feature selection, and it has strong adaptability to data sets; RF can process both discrete data and continuous data, and data sets need not be standardized.

(6) The training speed is fast and can be applied to large-scale data sets.

(7) Because each tree can be generated independently and simultaneously, it is easy to parallelize.

(8) Because of the simplicity, high accuracy and strong anti-overfitting ability of RF, it is suitable as a benchmark model when facing nonlinear data.

GBDT (Gradient Boosting Decision Tree) is a representative algorithm of the boosting algorithm series [39], which is an iterative decision tree algorithm that consists of multiple decision trees. The conclusions of all trees are added up as the final answer. The idea of GBDT is to represent the loss function by square error, in which each regression tree learns the conclusions and residuals of all previous trees, and fits to a current residual regression tree. In the process of GBDT iteration, assuming that the strong learner obtained by previous iteration is the loss function. The goal of this iteration is to find a weak learner for the regression tree model that minimizes the loss in this round. In other words, the decision tree found in this iteration should make the loss function of samples as small as possible. The advantage of GBDT is that it is suitable for low-dimensional data, and it can process nonlinear data. The disadvantage is that it is difficult to parallelize because of the connection between weak classifiers, and it will increase the computational complexity of the algorithm if the data dimension is high.

The eXtreme Gradient Boosting (XGBoost) method is an upgraded version of GBDT, in which the objective function is:

$$\zeta\left(\phi\right)=\sum_{\mathrm{i}} l\left(\hat{y}_i,y_i\right)+\sum\nolimits_{\mathrm{k}}\Omega\left(f_k\right)$$
$$\text{where}\quad \Omega\left(f\right)=\gamma T+\frac{1}{2}\lambda\left\|w\right\|^2 \tag{1}$$

K represents the total number of trees, f_k represents the kth tree, $l\left(\hat{y}_i,y_i\right)$ represents the sample x_i training error, $\hat{y}_i$ represents the predicted result of x_i, y_i represents the true value, T is the number of leaf nodes of the tree, γ and λ are Tunable parameter

The model predicted by the $t-$ th iteration is as follows:

$$\zeta^{(\mathrm{t})}=\sum_{i=1}^{n} l\left(y_i,\hat{y}^{(t-1)}+f_t\left(x_i\right)\right)+\Omega\left(f_t\right) \tag{2}$$

The error function is expanded by second-order Taylor expansion, and the formula is as follows:

$$\zeta^{(t)}\cong\sum_{i=1}^{n}\left[l\left(y_i,\hat{y}^{(t-1)}\right)+g_if_t\left(x_i\right)+\frac{1}{2}h_if_t^2\left(x_i\right)\right]+\Omega\left(f_t\right)$$
$$\text{where}\quad \mathrm{g_i}=\delta_{\hat{y}^{(t-1)}}l\left(y_i,\hat{y}^{(t-1)}\right)\ \mathit{and}\ h_i=\delta^2_{\hat{y}^{(t-1)}}l\left(y_i,\hat{y}^{(t-1)}\right) \tag{3}$$

Remove the constant term:

$$\zeta^{(t)}=\sum_{i=1}^{n}\left[g_if_t\left(x_i\right)+\frac{1}{2}h_if_t^2\left(x_i\right)\right]+\Omega\left(f_t\right) \tag{4}$$

The solution of the final error function is:

$$\bar{\zeta}^{(\mathrm{t})}\left(q\right)=-\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in g_i}g_i\right)^2}{\sum_{i\in h_i}h_i+\lambda}+\gamma T \tag{5}$$

$$w_j^* = -\frac{\sum_{i \in g_i} g_i}{\sum_{i \in h_i} h_i + \lambda} \tag{6}$$

XGBoost and GBDT are both gradient lifting and integrated learning, but they are different. The advantages of XGBoost over GBDT are as following:

(1) GBDT's base classifier only supports CART trees, while XGBoost supports linear classifiers, which is equivalent to logistic regression and linear regression with L1 and L2 regular terms.

(2) XGBoost adds a regular term to the objective function to control the complexity of the model. The regular term contains the number of leaf nodes and the modulus square of the fraction output on each leaf node. From the perspective of the trade-off between deviation and variance, the regularization term reduces the model variance, makes the learned model simpler and prevents overfitting.

(3) Traditional GBDT only uses the first derivative, whereas XGBoost expands the objective function with the second-order Taylor expansion, which is closer to the real loss function than the first-order Taylor expansion.

(4) To weaken the impact of each tree and give the subsequent trees more learning space, XGBoost multiplies the weight of leaf nodes into the learning rate, mainly after an iteration.

B. SVM

The basic idea of classification is to find a partitioning hyperplane in the sample space based on a training set $D = (x_1, y_1), (x_2, y_2), \ldots (x_m, y_m)$, where $y_i \in \{-1, +1\}$ that separates the different classes of the samples.

In the sample space, the division of hyperplanes can be described by the following linear equations:

$$W^{\mathrm{T}}x + b = 0 \tag{7}$$

Under linear separable conditions, assume that the data satisfies:

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x_i} + b) \geq +1 \tag{8}$$

Considering relaxation variables, the data needs to satisfy the following conditions:

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) + \varepsilon_i \geq +1 \tag{9}$$

Finally, the optimization equation is determined as follows:

$$\min_{\boldsymbol{w},b} \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{\mathrm{i}=1}^{N} \varepsilon_i \tag{10}$$

C is the model penalty parameters.

C. XGBSVM

The XGBSVM model combines the XGBoost and SVM algorithms, which uses the XGBoost method as a feature converter to construct a new feature combinations for training SVM model.

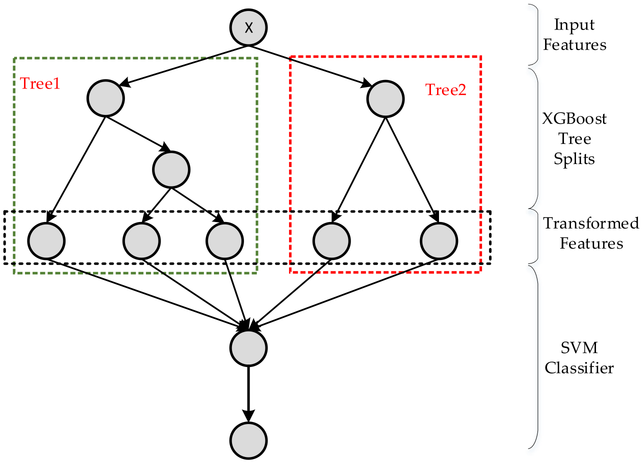

FIGURE 1. XGBSVM hybrid model structure.

The XGBoost method is an ensemble learning method that can generate many trees. The number of trees can be set according to actual problems. Assume that the XGBoost algorithm is used to construct two trees, Tree1 and Tree2, Each non-leaf node in Tree1 represents a judgment condition, and X, as an input sample, is judged according to the judgment condition of the non-leaf node in Tree1. After experiencing all the judgment conditions, it finally falls on a leaf node of Tree1. The leaf node where X falls is marked as 1, and the other leaf nodes are marked as 0, thus obtaining a sparse matrix of Tree1. Similarly, for Tree2, each non-leaf node is also a condition for judgment, sample X eventually falls on a leaf node according to the judgment condition, and the sparse matrix of Tree2 is obtained. Finally, the sparse matrix of Tree1 and Tree2 are concatenated to obtain the new features of X.

For example, there are two trees in Figure 1, the left tree has three leaf nodes, and the right tree has two leaf nodes. So the final feature is a five-dimensional vector. For input X, suppose it falls on the first node of Tree1, encoding (1, 0, 0), and on the second node of Tree2, encoding (0, 1), so the whole encoding is (1, 0, 0, 0, 1). This kind of encoding is input into SVM as a feature for classification. Because each tree path has its distinction, the characteristics and combinations of features obtained from the path are relatively distinctive, and the effect will not be inferior to manual experience in theory. The greatest advantage of this method is that it omits searching for features and feature combinations manually.

D. EVALUATION METRICS

In this paper, AUC (area under curve) and INE (improved normalized entropy) are used to evaluate the model.

AUC is a standard used to measure the quality of classification models. To solve the problem of unbalanced distribution in different categories, a new classification model performance judgment method, ROC (receiver operating characteristic) analysis, is introduced from the medical field. The main analysis tool of ROC is a curve drawn on a two-dimensional

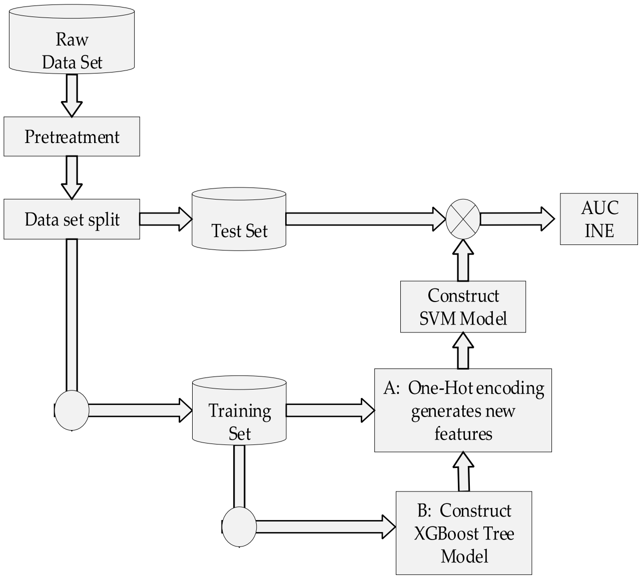

FIGURE 2. Constructing a prediction model flow chart.

plane, namely, the ROC curve. The horizontal coordinate is the false positive rate (FPR), and the vertical coordinate is the true positive rate (TPR). The value of AUC is the area below the ROC curve. Usually, the value of AUC is between 0.5 and 1.0, and larger AUC values represent better performance.

INE is improved on the basis of a paper published by Facebook in 2014, and it is suitable for measuring the model performance in this article [31]. INE equals the average logarithmic loss of each test set sample divided by the entropy of the test set.

In equation 11, N is the number of test sets, p_i is the predicted value, and P is the empirical probability of test samples, that is, the probability of label 1 in test samples. y_i is the label for the test sample, $y_i \in \{-1, 1\}$.

$$INE = \frac{-\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1+y_i}{2}\log(p_i) + \frac{1-y_i}{2}\log(1-p_i)\right)}{-(p*\log(p) + (1-p)*\log(1-p))} \tag{11}$$

In model evaluation, the smaller the INE value and the larger the AUC value are, the better the model performance is.

E. THE PROCESS OF CONSTRUCTING PREDICTION MODELS

The flow chart of the prediction model is shown in Figure 2. First, according to the physical examination data of hypertensive patients, in addition to selecting some basic human indicators, measurable indicators related to blood pressure and cardiovascular diseases are also selected. Next, the data are preprocessed, including filling missing values and normalization. The experimental data set is divided into two parts, the training set and the test set. The training set is also divided into training set A and training set B. Training set A is used to construct the tree model. In this paper, RF, GBDT and XGBoost are used to construct the model for comparative analysis. These three ensemble learning methods will construct many trees, so each sample in training set B will get its own position in the leaf node of each tree, such as the second leaf node of the second tree. Then, a new feature combination is obtained by using one-hot encoding. Combining the original label with the new features, a new training set is generated. Because it is a binary classification problem, SVM can be used, so the training set is put into the SVM model for training. To prove the high stability of the model, a cross-validation method is used. If the model is feasible, the probability of each test sample is obtained by inputting the test set into the model. Finally, AUC and INE are used to evaluate the model.

It should be noted that the classification criteria for each tree model node have been uniquely determined after training set A achieves the tree model by XGBoost, RF or GBDT. Therefore, training set B accurately locates each sample in the tree model, and each sample in training set B can be obtained by using the one-hot encoding method, the dimensions of new features are the same.

The data set in this article is from a hospital in Beijing, which contains patients from various provinces in China, and the total sample size is 1357.

The characteristics used in the model include basic patient indicators, such as height, weight; heart rate; limb blood pressure, including left upper limb systolic pressure (LARMSBP) and left upper limb diastolic pressure (LARMDSBP); 24-hour ambulatory blood pressure, including 24-hour average diastolic pressure(MEANDBP) and 24-hour average systolic pressure (MEANSBP); inflammatory factors, such as the erythrocyte sedimentation rate (ESR) and C-reactive protein (CRP); ultrasound electrocardiogram indicators, including LA, RAD; regular indicators, including white and red blood cell counts (WBC, RBC), hemoglobin (HB) ; and so on. In addition to the above indicators, urinary routine indicators, blood biochemical indicators, thyroid function indicators, urinary protein indicators, and cardiovascular related laboratory indicators were also used. The total number of indicators was 83. Table 1 shows all the features.

On the basis of the original features, the samples with missing values are deleted, the final data set contains 372 samples, 94 of which were identified as having hypertensive heart disease. The samples obtained after maximum and minimum normalization are shown in Table 2.

Because the data used in this paper are based on whether hypertensive patients have hypertensive heart disease in three years, they are scarce, and the sample size is only 372. According to the methods of this paper, the data should be divided into the test set and training set. At the meantime, one part of the training set is used to construct the tree model, the other part constructs new features through the tree model structure already constructed, so it is very important to divide the proportions of the original data set reasonably

TABLE 1. Original features.

Num	Name	Num	Name	Num	Name
1	SEX	29	USG1	57	RARMSBP
2	AGE	30	ALT	58	RARMDBP
3	HEIGHT	31	AST	59	LARMSBP
4	WEIGHT	32	K	60	LARMDBP
5	BMI	33	Na	61	RLEGSBP
6	HR	34	Cl	62	RLEGDBP
7	PULSE	35	GLU	63	LLEGSBP
8	RYSBPL	36	CREA	64	LLEGDBP
9	RYDBPL	37	BUN	65	BAPWVR
10	HTBEGIN	38	URIC	66	BAPWVL
11	ZGSBP	39	HSCRP	67	ABIR
12	ZGDBP	40	TG	68	ABIL
13	PSSBP1	41	TC	69	AHI
14	PSDBP1	42	HDLC	70	APNEA
15	AO	43	LDLC	71	HYPOPNEA
16	LA	44	FT3	72	SAO2
17	IVSD	45	FT4	73	MEANSAO2
18	LV	46	T3	74	MEANSBP
19	EF	47	T4	75	MEANDBP
20	LVPWd	48	TSH	76	HIGHSBP
21	RVd	49	MAUCR	77	HIGHDBP
22	WBC	50	HUPRO	78	LOWSBP
23	NEUT	51	HBLAC	79	LOWDBP
24	RBC	52	HCY	80	DAYMSBP
25	HB	53	ESR	81	DAYMDBP
26	PLT	54	CRP	82	NIHTMSBP
27	UKET	55	NTPRO	83	NIHTMDBP
28	USG	56	ET		

TABLE 2. Normalizes data.

NO.	BMI	HR		NIHTMSBP	NIHTMDBP
1	0.14	1.00	……	0.44	0.35
2	0.49	0.40	……	0.39	0.41
3	0.14	0.35	……	0.51	0.39
4	0.32	0.43	……	0.70	0.58
…	…	…	…..	…	…
…	…	…	…...	…	…
369	0.26	0.43	……	0.52	0.36
370	0.32	0.31	……	0.30	0.33
371	0.34	0.32	……	0.15	0.30
372	0.27	0.24	……	0.40	0.46

TABLE 3. Combination type and description.

Combination type	Description	$\frac{\text{Test set ratio}}{\text{Training set B ratio}}$
A	The test set accounts for 30% of the original data set, and training set B accounts for 50% of the training set	$\frac{0.3}{0.5}$
B	The test set accounts for 30% of the original data set, and training set B accounts for 40% of the training set	$\frac{0.3}{0.4}$
C	The test set accounts for 40% of the original data set, and training set B accounts for 40% of the training set	$\frac{0.4}{0.4}$
D	The test set accounts for 40% of the original data set, and training set B accounts for 50% of the training set.	$\frac{0.4}{0.5}$

and effectively. Therefore, four combinations are selected to analyze the influence of different partitions on the model. Table 3 is the type of combination and the corresponding description.

III. RESULTS AND DISCUSSION

To clearly describe the constructed tree structure, Figure 3 shows the second tree constructed by the XGBoost method in training set A, which has eight leaf nodes. The blue box represents the feature, and the f21 represents the 22nd feature of the original feature, because the feature count in the tree model starts from 0, and the number in the yellow box represents the weight of the leaf node. The next step is to traverse all the samples in training set B through all the trees in each model, and then determine the position of each sample on the leaf node of each tree. Finally, one-hot encoding is used, that is, if a sample falls on the second leaf node of the second tree, then the sample is coded as $x_1 = (0, 1, 0, 0, 0, 0, 0, 0,)$ on the second tree. Similarly, the sample traverses all the trees and gets the coding on each tree. Finally, by splicing all the

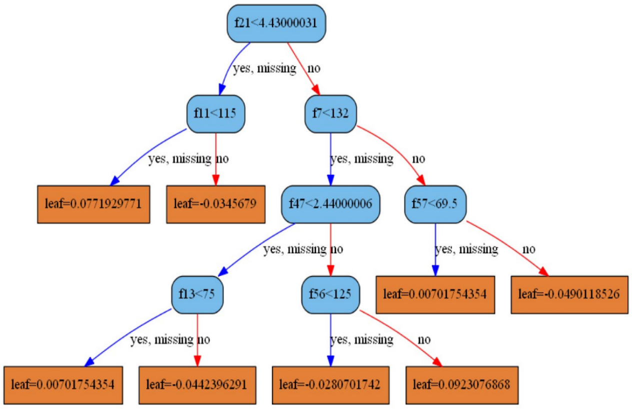

FIGURE 3. The second tree of XGBoost.

TABLE 4. Five-fold cross-validation results.

Method	Mean score
GBDT	81.50%
RF	78.80%
XGBoost	86.30%
RF+SVM	79.80%
GBDT+SVM	89.60%
XGBSVM	91.70%

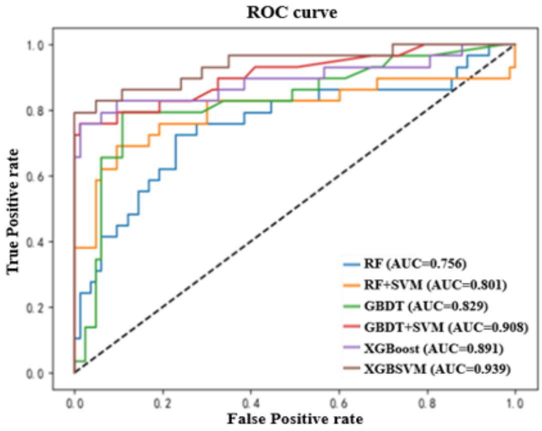

FIGURE 4. ROC curve of combination A.

coding together, a new feature combination for the sample is obtained, $X = (x_1, x_2, \ldots x_n)$, n is the number of trees. Using this method, a new feature combination for all samples in training set B can be obtained.

To verify the high stability and feasibility of this method, a five-fold cross-validation is carried out. Five methods are selected for comparative analysis to verify the excellent performance of XGBSVM. In the experiment, by changing the number of evaluators and the depth of the tree, the best combination of parameters is determined. For XGBoost, the number of evaluators is 40, and the depth of the tree is 3. The cross-validation results for these methods are shown in Table 4, which show that these methods are feasible and have high performance and can be used to predict test sets.

After combining SVM with RF, GBDT and XGBoost to form a hybrid model, ROC curves are shown in Figures 4-7 under the four combinations of A, B, C and D. Because of the small sample size, the ROC curve is not smooth. The abscissa of the ROC curve is the false positive rate (FPR), and the ordinate is the true positive rate (TPR). The TPR is the ratio of all actual positive samples correctly judged to be positive. The FPR is in the actual number of all negative samples wrongly judged as positive ratios. For a classifier, we can obtain a TPR and FPR point pair according to their performance on the test sample. In this way, the classifier can be mapped to a point on the ROC plane. By adjusting the threshold used in classifier classification, a curve passing through (0, 0), (1, 1) can be obtained, which is the ROC curve of the classifier. In general, the curve should be above the (0, 0) and (1, 1) lines.

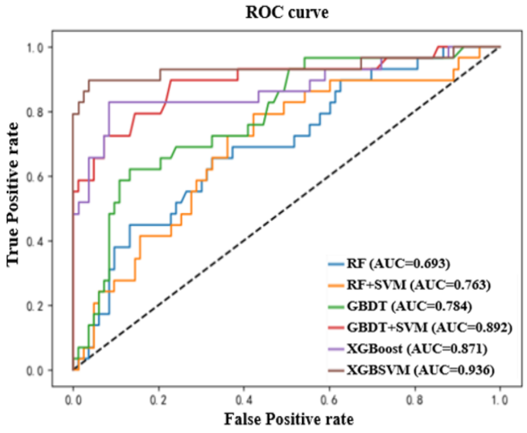

FIGURE 5. ROC curve of combination B.

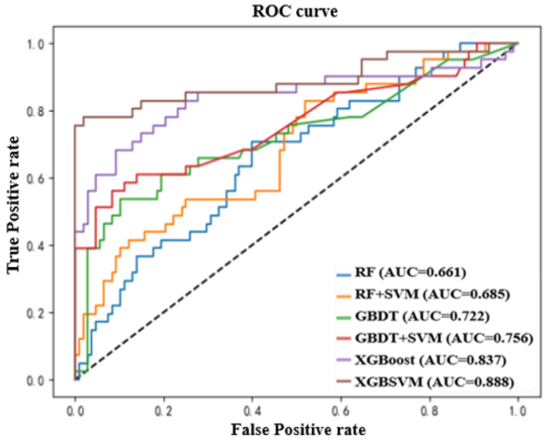

FIGURE 6. ROC curve of combination C.

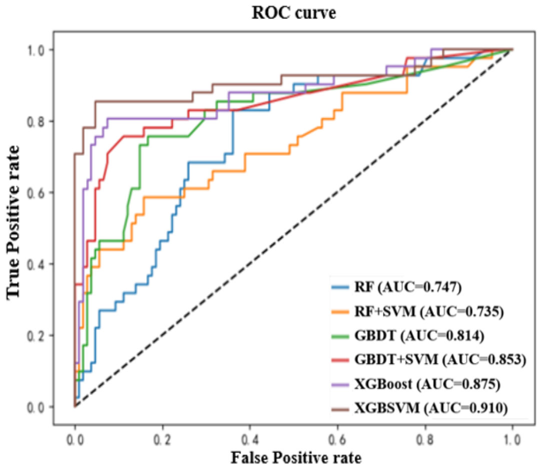

FIGURE 7. ROC curve of combination D.

After four combinations of ROC curves are obtained, the corresponding AUC value can be calculated. Similarly, different combinations of INE can be calculated separately. The AUC and INE of the six methods are shown in Table 5 and Table 6. The proportion of the test set refers

TABLE 5. AUC of different methods under different combinations.

Test set ratio	0.3	0.3	0.4	0.4
Training set B ratio	0.5	0.4	0.4	0.5
Evaluation Criterion	AUC			
GBDT	0.829	0.784	0.722	0.814
RF	0.756	0.693	0.661	0.747
XGBoost	0.891	0.871	0.837	0.875
RF+SVM	0.801	0.763	0.685	0.735
GBDT+SVM	0.908	0.892	0.756	0.853
XGBSVM	0.939	0.936	0.888	0.910

TABLE 6. INE of different methods under different combinations.

Test set ratio	0.3	0.3	0.4	0.4
Training set B ratio	0.5	0.4	0.4	0.5
Evaluation Criterion	INE			
GBDT	1.927	1.995	2.071	1.96
RF	2.045	2.17	2.266	2.2
XGBoost	1.369	1.525	1.66	1.668
RF+SVM	1.736	2.167	2.204	2.167
GBDT+SVM	1	1.565	1.81	1.583
XGBSVM	0.895	0.868	1.263	1.189

to the proportion in the original data set, and the proportion of training set B refers to the proportion in the training set. Considering RF, GBDT and XGBoost separately, the AUC value of XGBoost is the largest in four cases, and INE is the smallest. After RF, GBDT and XGBoost are combined with SVM to form a hybrid model, the AUC of the XGBSVM method is the largest and the INE is the smallest. Comparing XGBoost with XGBSVM, for AUC, XGBSVM is larger than XGBoost, and for INE, XGBSVM is smaller than XGBoost in the four cases. For a more intuitive analysis of the data in Table 5 and Table 6, AUC and INE are shown in Figures 8-10.

Figure 8 shows the AUC values of different models under the four combinations of A, B, C and D. According to the figure, the AUC values of each model are above 0.6. The maximum value is 0.939, corresponding to the XGBSVM model, and the minimum is 0.661, corresponding to the RF model. The performance of the hybrid model is better than that of a single model.

Figure 9 shows the INE values of different models under the four combinations of A, B, C and D. The smaller the INE, the better the model performance. According to the graph, the difference between different models and different combinations is large. The maximum value is 2.266, corresponding to the RF model, and the minimum is 0.868, corresponding to the XGBSVM model.

To better show the difference of INE in the different models under the different combinations, the maximum INE under the four combinations is taken as the standard, that is, 100%. The ratio of INE of each method to the worst model is calculated separately, as shown in Figure 10. In each combination, the XGBSVM hybrid model accounts for a small proportion, so the model performance is far superior to the RF. The GBDT + SVM hybrid model also has good performance, but compared with XGBSVM, the performance is slightly inferior.

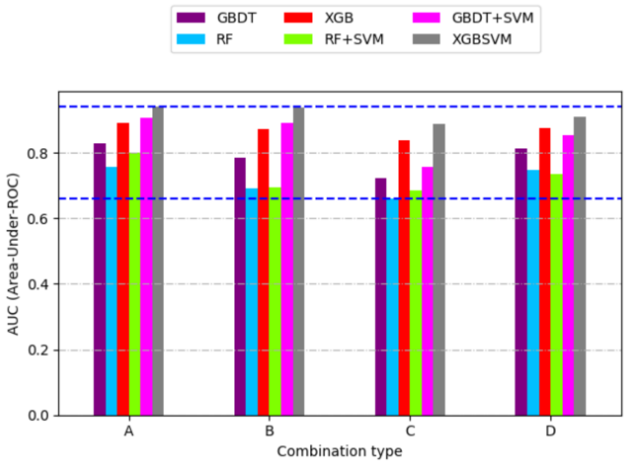

FIGURE 8. AUC of different combinations.

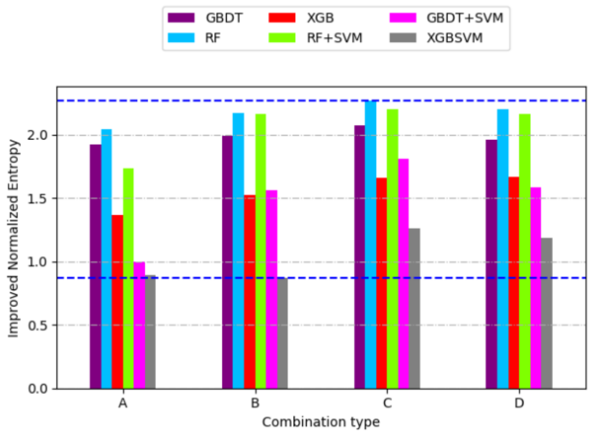

FIGURE 9. INE of different combinations.

It can be found that, in the four cases, the effect of A is the best because the number of samples in this dataset is relatively small; if the proportion of test sets is relatively small, then the proportion of training sets is larger, and more samples are used to train the model. Similarly, if the proportion of training samples A is large, more samples are used to construct tree models. The more samples there are, the better the performance of the tree models is. Therefore, among the six methods, combination A is the best method to segment samples.

Table 5 and Table 6 show that the XGBSVM and XGBoost models work better, consider these two models separately.

Figure 11 shows the impact of the number of trees on the INE and AUC of the two models. It can be seen from

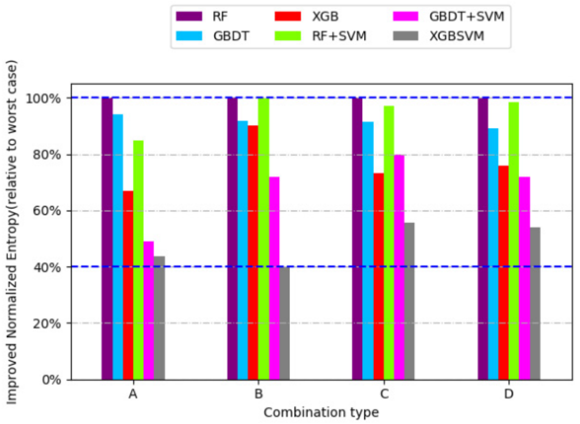

FIGURE 10. Based on maximum entropy, the proportion of INE in each model.

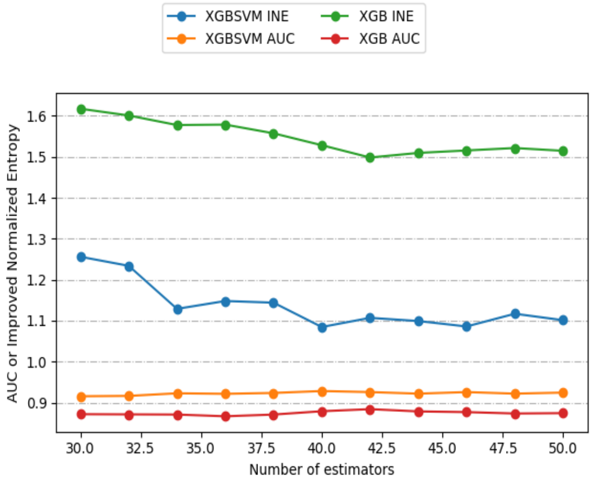

FIGURE 11. The influence of the number of trees.

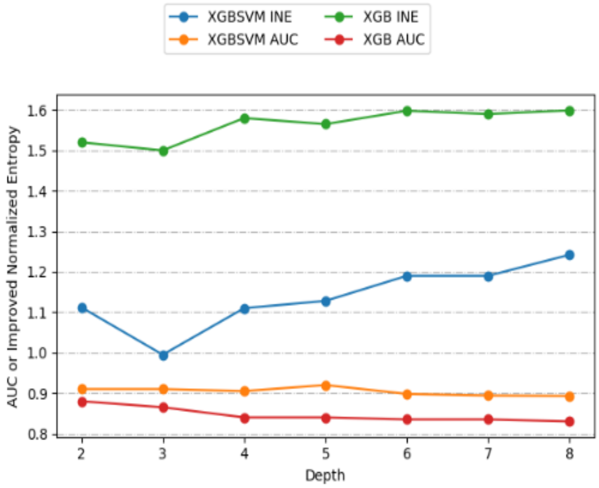

FIGURE 12. The influence of tree depth.

the graph that the AUC values of the two models basically remain unchanged with an increasing number of trees, but the INE of the two models decreases, and the INE of XGBSVM decreases more than the XGBoost. Comparing the two models, the INE of XGBSVM is 36.26% smaller than that of XGBoost, and the AUC is 5.51% higher.

Figure 12 shows the effect of tree depth on the INE and AUC of the two models. As seen from the graph, the AUC of the two models changes with increasing tree number. Among these models, the AUC of XGBoost decreases and to less than 0.9. Although the AUC of XGBSVM changed slightly, it remains stable at approximately 0.9. The INE of the two models reaches the minimum when the tree depth is 3. With increasing tree depth, the INE increases. The deeper the tree is, the more complex the model is, but the number of samples in this paper is relatively small. Therefore, when the tree depth reaches 3, a good model has been constructed. With increasing tree depth, the model may exhibit an overfitting phenomenon that results in worse model performance for the test samples. Combined with the four combinations, the INE of XGBSVM is 38.10% smaller than that of XGBoost, and the AUC is 7.44% higher. Through comparative analysis, XGBSVM has good performance and stability, and it can be used to construct prediction models.

IV. CONCLUSION

This paper presents a new hybrid XGBSVM model. To validate performance, the model was used to predict whether hypertensive patients develop hypertensive heart disease within three years. There are two evaluation indicators; one indicator is the traditional AUC, and the other indicator is the INE proposed in this paper. The XGBSVM is compared with five other models. The final experimental results prove that XGBSVM is highly feasible, stable and accurate, and it is a new model that can be applied practically. In the experiment, the final output of this model is the probability of patients being classified into positive categories. That is, the probability of patients suffering from hypertensive heart disease is the output. Doctors make their own judgment according to the actual situation. For example, when the threshold is set to 0.6, patients whose output probability is higher than 0.6 are considered to have hypertensive heart disease. When the threshold is set to 0.7, patients whose output probability is higher than 0.7 are all considered to have hypertensive heart disease. Therefore, the model is flexible and can make different judgments according to the actual situation.

For a new hypertensive patient, the probability of hypertensive heart disease within three years can be obtained by directly importing the required physical examination indicators into the established model. Over time, patients accumulate to a certain number and can be added to the original sample, and then the original sample is updated, so the tree model can be reconstructed. With increasing sample size, the model performance will be better, and the final output will be more reliable. In China, or around the world, with continuous improvement of living standards and the changes in people's lifestyles, increasing numbers of people are suffering from hypertension. The heart is one of the target organs of hypertension, so the number of people suffering from hypertensive heart disease will continue to increase. If it can be predicted in advance that patients with hypertension have a greater probability of developing hypertensive heart disease within three years, then doctors can recommend preventive treatment for these patients, which can not only greatly reduce their physical and mental pain but also ensure their life safety. Therefore, this study has strong practical significance and value.

The main contribution of this paper is to propose a new XGBSVM hybrid model. This model is based on the original machine learning theory and is a new machine learning development. XGBSVM not only eliminates the complicated work of feature filtering but also has high performance and stability. At the same time, a new evaluation index for entropy models is proposed. This paper not only applies the combination of entropy and artificial intelligence to the field of biomedicine but also proposes an effective new model and evaluation index in the field of biomedicine. Therefore, this paper makes the important contribution of entropy and artificial intelligence to the field of biomedicine.

In recent years, the application of entropy has been deepening, and it has been widely used in biology and medicine. Entropy can be used to analyze biological evolution, memory, pathology and so on. If the above content can be combined with entropy to establish a model, it will undoubtedly play an extremely important role in exploring the common law between the living and the non-living worlds. Similarly, machine learning has been developing rapidly in recent years, and its application field has been expanding continuously with outstanding contributions to the development of science and technology. Therefore, the combination of machine learning with other fields is an inevitable trend. The combination of biomedicine and machine learning will certainly make great progress.

In this paper, the newly proposed XGBSVM model is applied to predict whether hypertensive heart disease occurs in hypertensive patients within 3 years. This prediction is a successful application of information entropy and machine learning in the biomedical field. With the continuous development of information entropy and machine learning, model fusion is a new development direction. The hybrid model will have the advantages of the original model and produce more accurate results, which will bring more beneficial results to human development. As the research progresses, increasing good models will be proposed. In this paper, XGBSVM is applied to the field of biomedicine. Future work will also try to apply the model to other fields, while improving its performance.

ACKNOWLEDGMENT

(Wenbing Chang, Xueyi Wu, and Yiyong Xiao contributed equally to this work.)

REFERENCES

[1] D. Hou, X. Ren, C. Wang, X. Diao, X. Hu, Y. Zhang, Q. Shen, and J. Chen, "A whole foxtail millet diet reduces blood pressure in subjects with mild hypertension," *J. Cereal Sci.*, vol. 84, pp. 13–19, Nov. 2018.

[2] V. L. Roger, D. M. Lloyd-Jones, E. J. Benjamin, J. D. Berry, W. B. Borden, D. M. Bravata, S. Dai, E. S. Ford, C. S. Fox, H. J. Fullerton, and A. S. Go, "Heart disease and stroke statistics–2012 update: A report from the American Heart Association," *Circulation*, vol. 125, no. 1, pp. e2–e220, 2012.

[3] B. Unal, K. Sözmen, H. Arik, G. Gerçeklioglu, D. U. Altun, H. Simsek, S. Doganay, Y. Demiral, Ö. Aslan, K. Bennett, M. O'Flaherty, S. Capewell, and J. Critchley, "Explaining the decline in coronary heart disease mortality in Turkey between 1995 and 2008," *BMC Public Health*, vol. 13, no. 1, 2013, Art. no. 1135.

[4] K. Sliwa, D. Ojji, K. Bachelier, M. Böhm, A. Damasceno, and S. Stewart, "Hypertension and hypertensive heart disease in African women," *Clin. Res. Cardiol.*, vol. 103, no. 7, pp. 515–523, 2014.

[5] P. S. Collaboration, "Age-specific relevance of usual blood pressure to vascular mortality: A meta-analysis of individual data for one million adults in 61 prospective studies," *Lancet*, vol. 360, no. 9349, pp. 1903–1913, 2012.

[6] P. M. Kearney, M. Whelton, K. Reynolds, P. Muntner, P. P. K. Whelton, and J. He, "Global burden of hypertension: Analysis of worldwide data," *Lancet*, vol. 365, no. 9455, pp. 217–223, 2005.

[7] G. Su, H. Cao, S Xu, Y Lu, X Shuai, Y. Sun, Y. Liao, and J. Li, "Left atrial enlargement in the early stage of hypertensive heart disease: A common but ignored condition," *J. Clin. Hypertension*, vol. 16, no. 3, pp. 192–197, 2014.

[8] W. Wang and D. Peng, "Prevalence and risk factors of hypertension in urban and rural residents in Chengdu: A cross-sectional survey," *Chin. J. Evidence-Based Med.*, vol. 14, no. 2, pp. 165–168, 2014.

[9] Y. Luo and J. Shi, "Prevalence of hypertension and its control among middle aged and elderly residents in a community," *Chin. Gen. Pract.*, vol. 16, no. 31, pp. 3015–3018, 2013.

[10] R. Used, "Lipid-related markers and cardiovascular disease prediction," *Jama*, vol. 307, no. 2499, p. 506, 2012.

[11] S. K. Kunutsor, J. E. Kootstra-Ros, R. T. Gansevoort, R. P. F. Dullaart, and S. J. L. Bakker, "Circulating gamma glutamyltransferase and prediction of cardiovascular disease," *Atherosclerosis*, vol. 238, no. 2, pp. 356–364, 2015.

[12] M. Nagata, Y. Kiyohara, Y. Murakami, F. Irie, T. Sairenchi, K. Miura, T. Okamura, H. Ueshima, and T. Ninomiya, "Prediction of cardiovascular disease mortality by proteinuria and reduced kidney function: Pooled analysis of 39,000 individuals from 7 cohort studies in Japan," *Amer. J. Epidemiol.*, vol. 178, no. 1, pp. 1–11, 2013.

[13] P. T. Skretteberg, S. E. Kjeldsen, S. E. Kjeldsen, J. E. Erikssen, L. Sandvik, K. Liestøl, G. Erikssen, T. R. Pedersen, J. Bodegard, and I. Grundvold, "HDL-cholesterol and prediction of coronary heart disease: Modified by physical fitness?: A 28-year follow-up of apparently healthy men," *Atherosclerosis*, vol. 220, no. 1, pp. 250–256, 2012.

[14] Q. Du, Z. Wang, and K. Nie, "Application of entropy-based attribute reduction and an artificial neural network in medicine: A case study of estimating medical care costs associated with myocardial infarction," *Entropy*, vol. 16, no. 9, pp. 4788–4800, 2014.

[15] S. Zhu and H. Zhao, "Application of mutual information based on weighted entropy in medical image registration," *Comput. Eng. Appl.*, vol. 47, no. 16, pp. 207–210, 2011.

[16] N.-C. Hsieh, C.-C. Shih, H.-C. Keh, C.-H. Chan, and L.-P. Hung, "Intelligent postoperative morbidity prediction of heart disease using artificial intelligence techniques," *J. Med. Syst.*, vol. 36, no. 3, pp. 1809–1820, 2012.

[17] A. D. Dolatabadi, S. E. Z. Khadem, and B. M. Asl, "Automated diagnosis of coronary artery disease (CAD) patients using optimized SVM," *Comput. Methods Programs Biomed.*, vol. 138, pp. 117–126, Jan. 2017.

[18] P. Sinha, "Comparative study of chronic kidney disease prediction using KNN and SVM," *Int. J. Eng. Res. Technol.*, vol. 4, no. 12, pp. 608–612, 2015.

[19] K. Kaur and M. Singh, "Heart disease prediction system using ANOVA, PCA and SVM classification," *Int. J. Adv. Res., Ideas Innov. Technol.*, vol. 2, no. 3, pp. 1–6, 2016.

[20] Y. Ren, H. Fei, X. Liang, D. Ji, and M. Cheng, "A hybrid neural network model for predicting kidney disease in hypertension patients based on electronic health records," *BMC Med. Inform. Decis. Making*, vol. 19, no. 2, p. 51, 2019.

[21] J. Yu, Z. Xuan, X. Feng, Q. Zou, and L. Wang, "A novel collaborative filtering model for LncRNA-disease association prediction based on the Naïve Bayesian classifier," *BMC Bioinf.*, vol. 2, no. 1, p. 396, 2019.

[22] S. Vijayarani and S. Dhayanand, "Liver disease prediction using SVM and Naïve Bayes algorithms," *Int. J. Sci., Eng. Technol. Res.*, vol. 4, no. 4, pp. 816–820, 2015.

[23] R. Shinde, P. Patil, J. Waghmare, and S. Arjun, "An intelligent heart disease prediction system using k-means clustering and Naïve Bayes algorithm," *Int. J. Comput. Sci. Inf. Technol.*, vol. 6, no. 1, pp. 637–639, 2015.

[24] H. Greenspan, B. V. Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1153–1159, Mar. 2016.

[25] E. Decencière, G. Cazuguel, X. Zhang, G. Thibault, J.-C. Klein, F. Meyer, B. Marcotegui, G. Quellec, M. Lamard, R. Danno, D. Elie, P. Massin, Z. Viktor, A. Erginay, B. Laÿ, and A. Chabouis, "TeleOphta: Machine learning and image processing methods for teleophthalmology," *IRBM*, vol. 34, no. 2, pp. 196–203, 2013.

[26] D. Kermany and M. Goldbaum, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[27] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," present at the 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, San Francisco, CA, USA, Aug. 2016.

[28] H. Zheng, J. Yuan, and L. Chen, "Short-term load forecasting using EMD-LSTM neural networks with a XGBoost algorithm for feature importance evaluation," *Energies*, vol. 10, no. 8, p. 1168, Aug. 2017.

[29] L. Torlay, B. Perrone-Bertolotti, E. Thomas, and M. Baciu, "Machine learning–XGBoost analysis of language networks to classify patients with epilepsy," *Brain Inform.*, vol. 4, no. 3, pp. 159–169, 2017.

[30] W. Chen, K. Fu, J. Zuo, X. Zheng, T. Huang, and W. Ren, "Radar emitter classification for large data set based on weighted-XGBoost," *IET Radar, Sonar Navigat.*, vol. 11, no. 8, pp. 1203–1207, Aug. 2017.

[31] Y. Chen and Z. Tang, "Research on stock price prediction based on XGBoost algorithm with pearson optimization," *Inf. Technol.*, vol. 9, pp. 84–89, 2018.

[32] Y. Li, Y.-L. Zhou, X.-Z. Han, and Z. Wang, "The improvement and application of XGBoost method based on the Bayesian optimization," *J. Guangdong Univ. Technol.*, vol. 35, pp. 23–28, 2018.

[33] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, "Pegasos: Primal estimated sub-gradient solver for SVM," *Math. Program.*, vol. 127, no. 1, pp. 3–30, Mar. 2011.

[34] E. Pasolli, F. Melgani, D. Tuia, F. Pacifici, and W. J. Emery, "SVM active learning approach for image classification using spatial information," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 4, pp. 2217–2233, Apr. 2014.

[35] M. Lapin, M. Hein, and B. Schiele, "Learning using privileged information: SVM+ and weighted SVM," *Neural Netw.*, vol. 53, pp. 95–108, May 2014.

[36] Y. Yin, D. Xu, X. Wang, and M. Bai, "Online state-based structured SVM combined with incremental PCA for robust visual tracking," *IEEE Trans. Cybern.*, vol. 45, no. 9, pp. 1988–2000, Sep. 2015.

[37] W. G. Touw, J. R. Bayjanov, L. Overmars, L. Backus, J. Boekhorst, M. Wels, and S. A. F. T. van Hijum, "Data mining in the Life Sciences with Random Forest: A walk in the park or lost in the jungle?" *Briefings Bioinf.*, vol. 14, no. 3, pp. 315–326, 2013.

[38] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, "An assessment of the effectiveness of a random forest classifier for land-cover classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 67, pp. 93–104, Jan. 2012.

[39] A. Sakhnovich, "On the GBDT version of the Bäcklund-Darboux transformation and its applications to linear and nonlinear equations and Weyl theory," *Math. Model. Natural Phenomena*, vol. 5, no. 4, pp. 340–389, 2010.

[40] X. He, O. Jin, T. Xu, B. Liu, T. Xu, Y. Shi, A. Atallah, R. Herbrich, S. Bowers, J. Q. Candela, and J. Pan, "Practical lessons from predicting clicks on ads at Facebook," in *Proc. 8th Int. Workshop Data Mining Online Advertising*, New York, NY, USA, Aug. 2014, pp. 1–9.

WENBING CHANG received the Doctor of System Engineering degree from Beihang University, in 2012. He is currently an Associate Professor with Beihang University. His research interests include system safety, risk management, and economic affordability.

YINGLAI LIU received the bachelor's degree in industrial engineering from the University of Electronic Science and Technology of China, in 2018. He is currently pursuing the master's degree in control science and engineering with Beihang University. His research interests include data analysis and data mining.

XUEYI WU received the doctor's degree in clinical medicine from the Peking Union Medical College, in 2012. She is currently with the Department of Cardiology of Fuwai Hospital, Chinese Academy of Medical Sciences, and the Peking Union Medical College. Her research interests include clinical and basic research of hypertension and other cardiovascular diseases.

YIYONG XIAO received the doctor degree in system engineering from Beihang University, in 2003. He is currently an Associate Professor with Beihang University. His research interests include economic affordability, data mining, network security optimization, and algorithm research.

SHENGHAN ZHOU received the Doctor of Management Science degree from Beihang University, in 2009. He is currently a Lecturer with Beihang University. His research interests include system safety, data mining, and risk management.

WEN CAO received the B.E. degree in safety engineering from the Capital University of Economics and Business, China, in 2009, and the M.S. and Ph.D. degrees in industrial and systems engineering from the State University of New York, Binghamton, USA, in 2011 and 2014, respectively. He has been a Performance Improvement Consultant with the Peninsula Regional Medical Center, since 2015. His research interests include modeling of healthcare delivery systems and human factors. He is a member of the Institute of Industrial Engineers (IIE) and Alpha Pi Mu Industrial Engineering Honor Society.