IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2019.DOI

Survival Risk Prediction of Esophageal Cancer Based on Self-organizing Maps Clustering and Support Vector Machine Ensembles

JUNWEI SUN[1], YULI YANG[1], YANFENG WANG[1], LIDONG WANG[2], XIN SONG[2], AND XUEKE ZHAO[2]

[1]School of Electrical and Information Engineering, Zhengzhou University of Light Industry, Zhengzhou, 450002, China
[2]State Key Laboratory of Esophageal Cancer Prevention & Treatment and Henan Key Laboratory for Esophageal Cancer Research of The First Affiliated Hospital, Zhengzhou University, Zhengzhou, 450052, China

Corresponding author: Yanfeng Wang (e-mail: yanfengwang@yeah.net).

ABSTRACT This article provides a method based on self-organizing maps (SOM) neural network clustering and support vector machine (SVM) ensembles to predict the survival risk levels of esophageal cancer. Nine blood indexes related to patient survival are found by using SOM clustering method. Two critical thresholds for survival are found by plotting the receiver operating characteristic (ROC) curve twice, and the lifetime is divided into three risk levels. Using the SVM method, patients' risk levels are predicted and assessed. Four kernel functions of SVM are compared, and the prediction effect of RBF kernel function is better than other kernel functions. The parameters of SVM are optimized by using genetic algorithm (GA), particle swarm algorithm (PSO) and artificial bee colony (ABC) algorithm. Experimental results show that the prediction accuracies are improved by using optimization algorithms. After comparison, ABC-SVM has better prediction results than GA-SVM and PSO-SVM with a high prediction rate and fast running time.

INDEX TERMS artificial bee colony, genetic algorithm, particle swarm optimization, self-organizing maps, support vector machine.

I. INTRODUCTION

ENORMOUS social and economic burdens have been caused by all kinds of tumors, which have been one of the leading causes of death in the whole world. With a high morbidity and mortality, esophageal cancer has become one of the leading causes of death worldwide, ranking sixth in the cause of making a deal of deaths every year [1]. Therefore, it is of significance for physicians to predict the survival risk of esophageal cancer.

Although the treatment methods and concepts of esophageal cancer have been improved gradually with the rapid development of science and the medical technology [2], the survival risk prediction of esophageal cancer still has some imperfections. Traditional statistical analysis is still widely used in the research on various pathological data of esophageal cancer patients. Because of the pathological complexity, some errors exist in the manual diagnosis. So a system that will can predict survival risk level is to be designed to solve the above problems. With the help of the system, the survival rates of esophageal cancer patients will be improved.

Recently, the method of machine learning is applied to the diagnosis of cancer more and more, due to the rapid development of computer-aided technology. Different algorithms have been explored to analyze the influencing factors of cancer and predict the risk level of cancer. Different statistical and machine learning techniques have been used to develop cancer prediction models, including random forest [3], extreme learning machine [4], naive bayes [5], artificial neural networks [6], and support vector machine [7].

To be specific, the genetic information is the base of most of the current cancer predictions [8]–[13]. Good prediction results have been achieved through these methods. However, the genetic information of patients must be obtained in the method. In this article, a new method based on blood index information of patients to predict the risk levels of esophageal

cancer is proposed. Self-organizing competitive maps (SOM) clustering and support vector machine (SVM) ensembles are used in the new method. A combination of multiple blood indicators which are closely associated with esophageal cancer survival is found, and a predictive system to predict the survival risk levels of esophageal cancer is established.

In this article, SOM neural network and SVM ensembles are used to find blood indexes that are significantly related to patient survival, and to predict patients' risk levels. At first, the SOM neural network is used to cluster seventeen blood indexes of patients, and a combination of nine indexes is found. It is verified that the combination of these nine indicators has a significant correlation to the survival of patients through the COX regression method of MedCalc software. Two critical thresholds for survival are attained by plotting the ROC curve twice and calculating the Youden index. Survival time is divided into three risk levels, where the patients' nine blood indexes and three risk levels are obtained. The patients' risk level is predicted and classified through the algorithm of SVM. Liner function, polynomial function, radial basis function (RBF) and sigmoid function, these four kernel functions are selected for SVM modeling. It is proved that the effect of the RBF kernel function is better than other kernel functions. The kernel parameters c and g of RBF are optimized so as to improve the performance of the SVM. The genetic algorithm (GA), particle swarm optimization (PSO), and artificial bee colony (ABC) algorithms are used to optimize the parameters, respectively. It is concluded that ABC is better than GA and PSO after comparing the predictions of the three optimization algorithms. The new method provided in this article for survival diagnosis of esophageal cancer can predict the survival risk levels of esophageal cancer accurately and effectively.

The purpose of this article is to study the survival risk prediction of esophageal cancer patients based on information of blood indicators. By using SOM clustering, ROC, SVM, GA, PSO, and ABC algorithm, a new method for predicting the survival risk of esophageal cancer is provided. Main contributions of this paper are summarized as follows:

(i) Based on the SOM clustering method, nine blood index combinations that are significantly related to the survival of esophageal cancer patients are found.
(ii) Three survival risk levels of esophageal cancer patients are gotten based on ROC method.
(iii) The risk level of patients with esophageal cancer is effectively predicted by ABC-SVM, based on the multiple indicator combination and three risk levels obtained above.

The extraction of relevant blood indicators is given in Section II. The Section III explains the division of survival risk levels. The survival risk prediction is given in Section IV. The optimization of SVM and the results discussion are described in Section V. Conclusions are made in Section VI.

II. DATA PROCESSING AND CORRELATION INDEX EXTRACTION

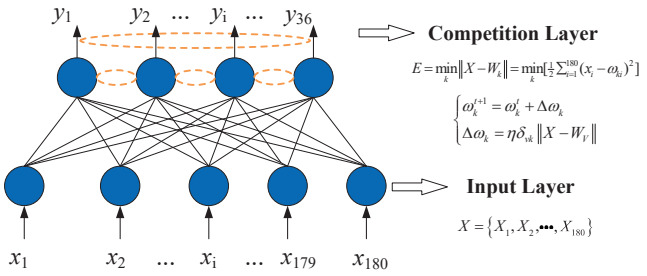


FIGURE 1. SOM neural network of 180-36 structure. There are 180 neurons in the input layer and 36 neurons in the mapping layer. X represents the input vector, i stands for the i-th node of the input layer, k is regarded as the k-th node of the output layer, t represents the number of network learning iterations, ω is considered as the connection weight value, and η is the learning rate.

A. EXTRACT MULTIPLE CORRELATION INDICATORS BASED ON SOM CLUSTERINGS

SOM neural network is an unsupervised learning neural network with self-organizing functions, which has the ability to map high dimensional inputs to low dimensions [14]. A two-layer network is composed of an input layer and a competition layer, and the neurons between the two layers implement a two-way connection [15].

As shown in Fig. 1, a self-organizing neural network of 180-36 structure is established. There are seven steps in the SOM clustering process:

Step1 Data selection and normalization

Seventeen blood indicators for 180 patients' information are selected, including WBC count, lymphocyte count, monocyte count, neutrophil count, eosinophil count, basophil count, red blood cell count, hemoglobin concentration, platelets count, total protein, albumin, globulin, PT, INR, APTT, TT, FIB. These 17 blood indexes are clustered and analyzed. The 180 patients' information of 17 blood indicators is normalized to $[-1, 1]$ by the $mapminmax$ function, and brought into the SOM model for clustering of blood indexes. The purpose of normalization is to make the algorithm converge quick and reduce error. The $mapminmax$ function is calculated by (1),

$$y = \frac{(y_{max} - y_{min})(x - x_{min})}{(x_{max} - x_{min})} + y_{min} \tag{1}$$

where y_{max} is 1 and y_{min} is -1.

Step2 Network initialization

Randomly set the vector of the initial connection weight value between the mapping layer and the input layer, $k \in [1, 36]$. The maximum number of learning cycles is given to 10, 50, 100, 200, 500, 1000, 2000, respectively. The initial value η of the set learning rate is 0.7, $\eta \in (0, 1)$. The initial neighborhood is set to N_{k0}.

Step3 Input of input vector

Input vector X that is input to the input layer can be expressed as (2).

$$X = (x_1, x_2, x_3, \cdots, x_m)^T, m \in [1, 17] \tag{2}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3007785, IEEE Access

Step4 Calculate the distance between the weight vector of the mapping layer and the input vector

The first set of training sample is randomly selected by (3),

$$X_i^l = \left(X_1^l, X_2^l, \cdots, X_i^l\right) \tag{3}$$

where i is the i-th node of the mapping layer, $i = 1, 2, ..., 180$, l is the training data, $l = 1, 2, ..., 17$. The closest neuron to the input vector will be found as the winning neuron depending on the size of the connection weight. The error function E is defined as the distance between the input vector and the connection weight vector. E is calculated by (4),

$$E = \min_k \|\boldsymbol{X} - \boldsymbol{W}_k\| = \min_k \left[\frac{1}{2}\sum_{i=1}^{180}(x_i - w_{ki})^2\right] \tag{4}$$

where k is the k-th node of the output layer, $k = 1, 2, ..., 36$, w_{ki} is the connection weight value of the i-th neuron of the input layer and the k-th input neuron of the mapping layer.

Step5 Weight learning

The weight of the winning neuron k is updated. The connection weights of the neurons around the winning neuron and the input vector are also updated, according to (5),

$$\begin{cases} w_k^{t+1} = w_k^t + \Delta w_k \\ \Delta w_k = \eta \delta_{vk} \|\boldsymbol{X} - \boldsymbol{W}_v\| \end{cases} \tag{5}$$

where t is the number of learning cycles, and $\boldsymbol{W}_v$ is the weight of the connection between the neurons surrounding the winning neurons and the input vector. η is a constant of $[0, 1]$, which gradually decreases to 0 by (6),

$$\eta(t) = 0.2\,(1 - t/1000) \tag{6}$$

δ_{vk} represents the value of the proximity relationship between the neuron k and the adjacent center v, as in (7),

$$\delta_{vk} = e^{-(D_{vk}/R)^2} \tag{7}$$

where D_{vk} represents the distance of the output neuron k from the center of the network topology to the adjacent center v. R is the radius of the winning neighborhood N_{kt} of neuron k.

Step6 Winning neurons are labeled. Returning to $Step3$, the next set of training data is selected. When 17 sets of training data are all completed, proceed to $Step7$.

Step7 $Step3$ is cycled to $Step6$. When the maximum number of learning cycles is reached, the loop ends.

The Big-O notation measures the worst-case complexity of an algorithm. In this study, the time complexity of the algorithm is evaluated by Big-O complexity. The execution number function of the SOM clustering algorithm is calculated by:

$$f(n) = a + 17 + 180 + n \tag{8}$$

a is a constant, represents the learning cycles. n represents the number of neurons in the mapping layer. Therefore, the time complexity of SOM mainly depends on the computational efficiency of the winning neurons. The time complexity of SOM clustering can be expressed as:

$$T_n = O(f(n)) = O(n) \tag{9}$$

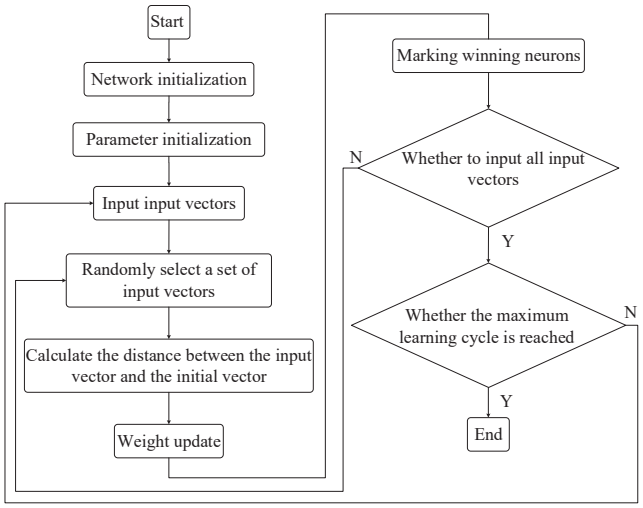


FIGURE 2. The SOM clustering flowchart

The SOM clustering flowchart is shown in Fig. 2. The results of using SOM clustering are shown in Table 1. Table 1 is the clustering result of SOM with learning cycles of 10, 50, 100, 200, 500, 1000, 2000 generations. Each row is the clustering result of seventeen blood indicators. The same type of blood indicators are represented by the same number. When the number of iterations of the SOM algorithm is 50, the blood indicators clustering effect is better. However, when the number of iterations is too large, resulting in unsatisfactory classification. The results of each iteration can also be verified by the subsequent COX regression. The smaller the P value is, the greater the correlation is. Therefore, the combination of blood indicators that has a greater correlation with the survival of the patients is 1, 2, 3, 4, 5, 6, 7, 13, 14, which corresponds to WBC count, lymphocyte count, monocyte count, neutrophil count, eosinophil count, basophil count, red blood cell count, PT, INR.

The SOM algorithm has certain limitations and shortcomings. First, the SOM algorithm needs to select an appropriate learning rate, and the size of the learning rate determines whether the performance of the SOM algorithm tends to be stable. Second, sometimes the initial weight vector of a neuron is too far from the input vector, which will cause it not to win the competition, and thus never learn and become useless neurons.

B. COX REGRESSION ANALYSIS TO VERIFY INDEX CORRELATION

The COX regression model is proposed by the British statistician D.R. Cox (1972). It can analyze the impact of multiple factors on survival. Due to the excellent performance, the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3007785, IEEE Access

TABLE 1. SOM clustering output.

Iteration steps	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}	c_{11}	c_{12}	c_{13}	c_{14}	c_{15}	c_{16}	c_{17}
10	13	1	1	1	1	1	1	24	36	30	3	3	5	1	3	3	36
50	1	1	1	1	1	1	1	24	36	27	6	9	1	1	6	2	36
100	26	31	31	32	31	31	25	18	35	5	3	15	20	1	9	13	36
200	13	1	1	7	1	1	3	22	24	33	31	20	14	1	26	4	36
500	33	32	31	33	31	31	19	18	29	5	9	1	13	1	3	21	36
1000	20	33	32	27	32	32	25	17	24	5	9	1	36	1	3	13	36
2000	19	25	31	26	31	31	33	18	29	5	16	1	13	1	3	14	36

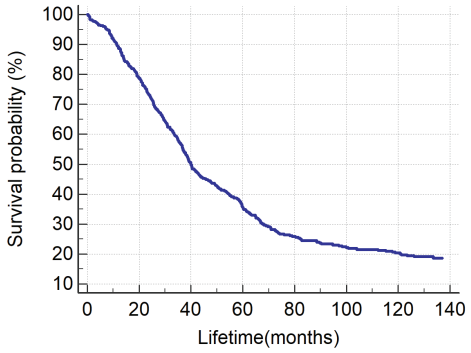


FIGURE 3. Survival function at the mean of the covariate. The survival months is taken as the time, the nine blood indicators obtained from the cluster are used as covariates.

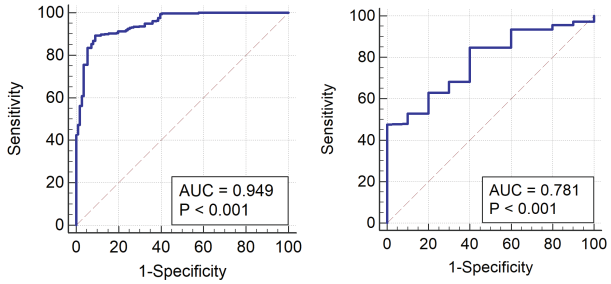


FIGURE 4. ROC curve analysis. (a) ROC curve of all samples. (b) ROC curve of samples with less than 27 months of survival. The ordinate is "Sensitivity" and the abscissa is "1-Specificity", the curve is clearly located at the upper left of the diagonal and has a good significance.

TABLE 2. Youden index.

Project	ROC for all samples	ROC for low survival samples
Youden index J	≤ 67.39	≤ 27.38
Sensitivity	89.24	47.65
Specificity	90.99	100

model has been widely used in the medical field since its proposal, and it is the most widely used multi-factor analysis method in survival analysis [16]–[18].

"MedCalc 18.2.1" software is used to make the COX model. The survival function at the mean of covariates is shown in Fig. 3. The results show that the P value of the overall score of the nine blood indicators is 0.0041 far less than 0.05. The combination of these nine blood indexes is significantly related to the patients' survival.

III. DIVIDE RISK LEVELS BASED ON ROC CURVE

The receiver operating characteristic (ROC) curve is also known as the sensitivity curve [19]–[22]. It can easily detect the ability to recognize performance at any threshold and select the best diagnostic threshold [23]–[26].

The ROC curve is used to determine the optimal cutoff values for the survival period. It is plotted with the survival month of all samples as the variable, named "ROC for all samples", as shown in Fig. 4(a). The value of area under curve (AUC) is 0.949, larger than 0.5, $P < 0.001$. It's obvious that a good threshold can be found for the second classification of survival. There is a better critical point which divides the lifetime into two levels. For the survival period, a critical threshold can be found to divide the survival period into two risk levels. The Youden index is calculated by using (10),

$$Youden\ Index = Sensitivity - (1 - Specificity) \quad (10)$$

survival value with the largest Youden Index is the critical threshold for survival time. Here, the threshold for survival is 67.39 months. The Youden index is shown in Table 2.

Then, in order to get two critical thresholds for survival, the sample information with a survival period of fewer months than 67.39 is summarized. Similarly, according to the above method, the ROC curve is drawn with the survival months as the variable, named "ROC for low survival samples", as shown in Fig. 4(b). The Youden index is shown in Table 2. The value of AUC is 0.781. The threshold for survival is 27.38 months by calculating the Youden index.

Therefore, the lifetime is divided into three risk levels. The survival period for not more than 27.38 months is seen as "risk level 1", the survival period for 27.38 to 67.39 months is seen as "risk level 2", and the survival period for more than 67.39 months is seen as "risk level 3", as shown in Fig. 5. Thus, the patients' characteristic data set is obtained, which contains nine blood indexes and three levels, as shown in Table 3.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3007785, IEEE Access

TABLE 3. Data of three risk levels of esophageal cancer patients.

WBC count	Lymphocyte count	Monocyte count	Neutrophil count	Eosinophil count	Basophil count	Red blood cell count	PT	INR	Life time (months)	Risk Level
5.7	1.7	0.3	3.3	0.3	0.1	4.61	9.1	0.66	13.3	1
7	1	0	5	0	0	4.6	11.7	0.94	0.26	1
5.8	1.5	0.3	3.9	0	0.1	4.6	7.8	0.53	16	1
.....										
6.6	1.3	0.4	4.7	0.1	0.1	4.91	8.1	0.56	62.6	2
7	2	0	4	0	0	4.9	8.7	0.62	28.7	2
6	2	0	3	0	0	4.9	11.2	0.89	65.2	2
.....										
7.5	2.6	0.6	4.2	0.1	0	3.72	7.1	0.46	82.7	3
6.4	1.2	0.6	4.5	0.1	0	3.72	9.2	0.67	134	3
7.1	1.5	0.4	4.8	0.1	0.3	3.63	8.8	0.63	77.8	3

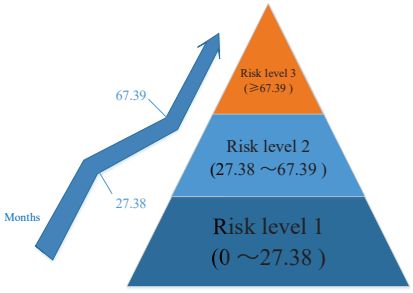


FIGURE 5. Risk levels.

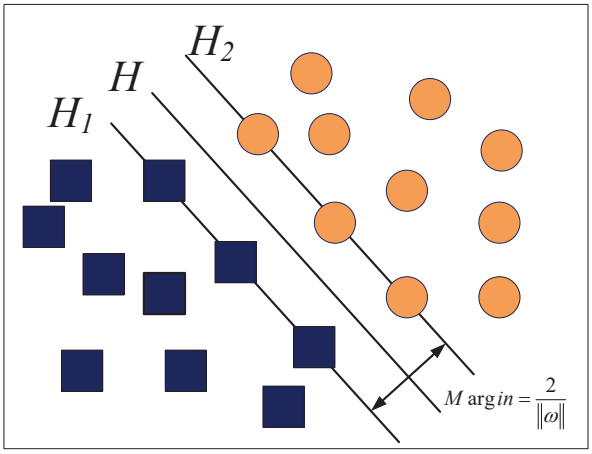


FIGURE 6. Support vector machine optimal classification surface.

IV. RISK LEVEL PREDICTION BASED ON SVM

SVM is a machine learning method based on the structural risk minimization criterion. It can solve high-dimensional problems and local minimum values, and its learning model has a good ability to promote [27]–[29].

As shown in Fig. 6, square and circular points represent two samples. H is an optimal classification line. H_1 and H_2 are the samples that are closest to the classification line and parallel to the classification line of each type. The point (x_i, y_i) is called the support vector. $Margin$ is the distance between H_1 and H_2. When extended to a high-dimensional space, the optimal classification line becomes the optimal classification plane. Support vectors are the data points closest to the decision plane and are the most difficult to classify data points, so they are directly related to the optimal position of the decision plane [30].

Choosing the SVM kernel function is an important step for survival risk levels prediction of esophageal cancer. The kernel function can map data from low-dimensional to high-dimensional, which can solve linear indivisible problems very well. Sigmoid function, polynomial function, linear function, and radial basis function are commonly used kernel functions of SVM.

The linear kernel function is expressed as:

$$K\left(x,z\right)=x^{T}z \tag{11}$$

The polynomial kernel function is decided by:

$$K\left(x,z\right)=\left(gx^{T}z+r\right)^{p},g>0 \tag{12}$$

The RBF is calculated by:

$$K\left(x,z\right)=exp\left(-g\left\|x-z\right\|^{2}\right),g>0 \tag{13}$$

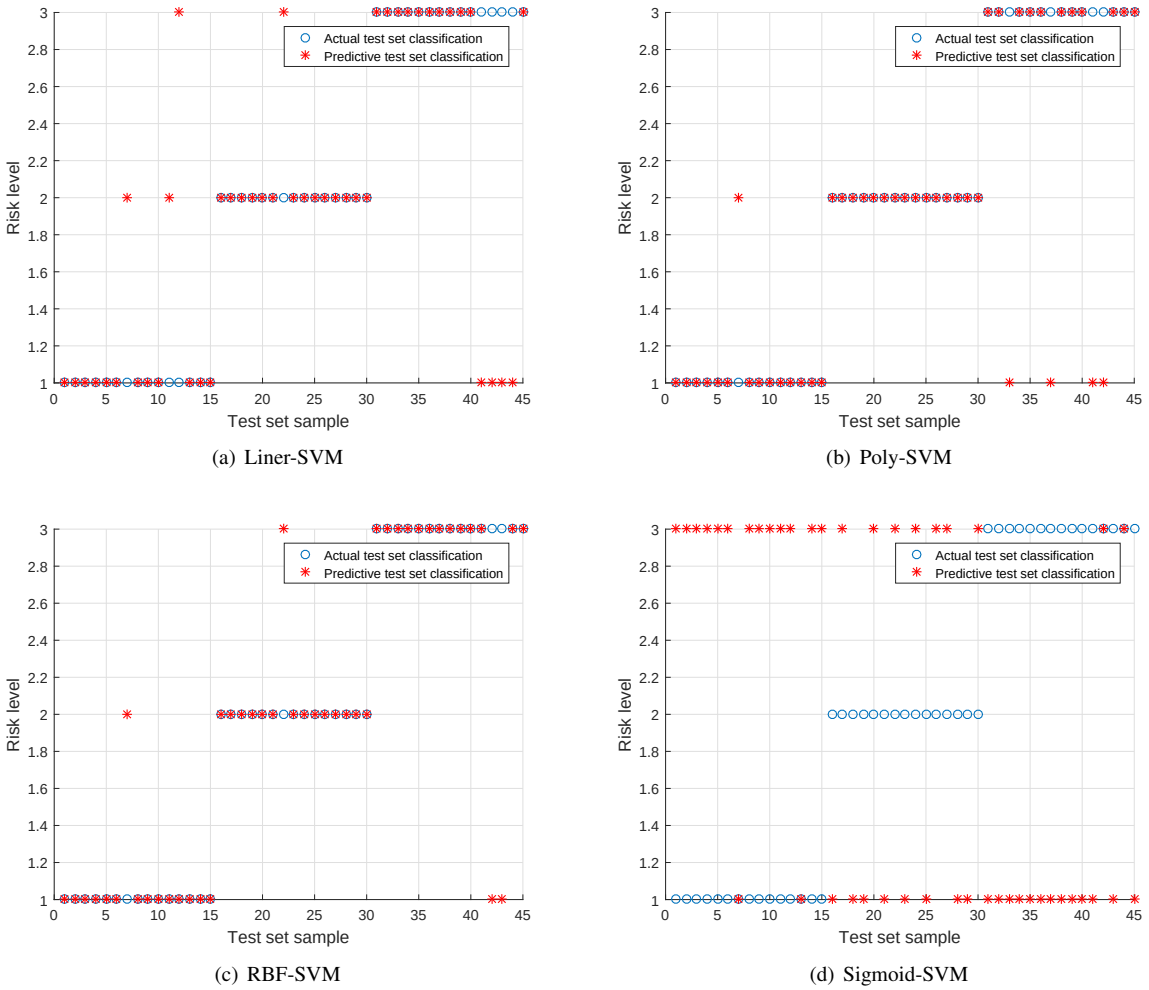
The sigmoid kernel function could be written as:

$$K\left(x,z\right)=tanh\left(gx^{T}z+r\right) \tag{14}$$

The parameter g is the parameter coefficient of the kernel function, which is the key to enhancing the performance of the SVM. The parameter r and parameter h are arbitrary constants. The parameter p is the power of the polynomial.

In the Windows 10 operating environment, MATLAB R2016a software is used to simulate methods of risk level prediction. All 180 esophageal cancer patients' information of three risk levels is investigated, and nine blood indicators are extracted from each patient's information using SOM neural network. Three risk levels of 135 cases of esophageal cancer cases are used as training samples, and 45 samples of each risk level are used for training. Three risk level data from 45 cases are used as test samples, and 15 samples of each risk level are used for testing. The normalization function $mapminmax$ is used to normalize the training sets and test sets. The data is normalized to the interval [-1,1].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3007785, IEEE Access



(a) Liner-SVM (b) Poly-SVM

(c) RBF-SVM (d) Sigmoid-SVM

FIGURE 7. SVM prediction results with different kernel functions.

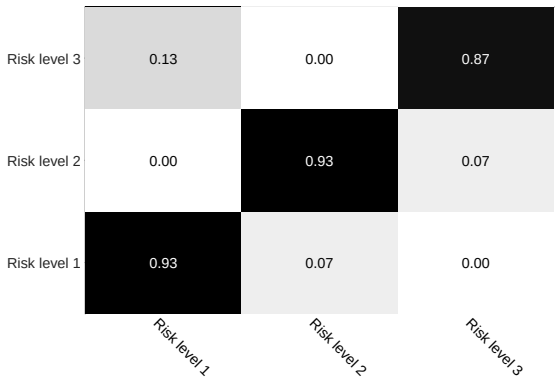


FIGURE 8. Confusion matrix of RBF-SVM.

The prediction results of the four kernel functions are shown in Fig. 7. The comparison of the prediction results of different kernel functions are shown in Table 4. The SVM has achieved good results in predicting the survival risk levels of esophageal cancer. The prediction accuracy rate of RBF-SVM without parameter optimization is 91.11%. It fully embodies the powerful classification of SVM algorithms and the unique advantages of classification and recognition in nonlinear and high dimensions.

The confusion matrix of RBF-SVM is shown in Fig. 8. The RBF-SVM predicts the risk levels of esophageal cancer very well. For the prediction results of 45 samples in the test set, 14 cases are predicted correctly for the risk level 1 with the accuracy rate of 0.93, 14 cases are predicted correctly for the risk level 2 with the accuracy rate of 0.93, and 13 cases are predicted correctly for the risk level 3 with the accuracy rate of 0.87. A total of 41 cases are predicted correctly in the test set with the accuracy rate of 91.11%.

V. SVM BASED ON PARAMETER OPTIMIZATION

The RBF kernel function needs to set two parameters, the penalty parameter c and the kernel parameter coefficient g, which also have an important effect on the result. In order to enhance the classification accuracy of SVM, the

TABLE 4. Comparison of prediction results of different kernel functions.

Kernel function	Risk levels	Correctly identify number	Recognition accuracy rate (%)	All correctly identify number	Overall recognition rate (%)
Liner	leve1	13	86.67	38	82.22
	leve2	14	93.33		
	leve3	11	73.33		
Polynomial	leve1	14	93.33	40	88.89
	leve2	15	100.0		
	leve3	11	73.33		
Radial basis function	leve1	14	93.33	**41**	**91.11**
	leve2	14	93.33		
	leve3	13	86.67		
Sigmoid	leve1	2	13.33	4	8.89
	leve2	0	0		
	leve3	2	13.33		

penalty parameter c and the kernel parameter coefficient g are optimized.

In recent years, meta-heuristic algorithms have been extensively studied, such as simulated annealing algorithm [31], tabu search algorithm, genetic algorithm, ant colony algorithm, particle swarm algorithm, artificial bee colony, fish swarm algorithm, cat swarm optimization, whale optimization algorithm, artificial algae algorithm, etc. The meta-heuristic algorithm provides a practical solution for complex optimization problems, which has been widely used in various fields and has achieved certain results. The study of synthetic polyurethane foam [32], simulated annealing algorithm is used to optimize model parameters. Artificial ant colony algorithm is used to solve the problem of optimizing the allocation of ship berths, which minimizes the time of container turnover at the terminal [33]. Combined tabu search algorithm and ant colony algorithm, the optimization performance is improved [34]. In 2016, Seyedali Mirjalili and Andrew Lewis proposed the whale optimization algorithm [35]. WOA can solve the problems of local optimal stagnation and slow convergence speed [36]. Meta-heuristic optimization algorithms have been recognized in machine learning, and they can find the best solutions to complex problems in science and engineering [37]. Different optimization algorithms have different special applications, and scholars need to find suitable meta-heuristic optimization algorithms in practical applications. In our study, GA, PSO and ABC are used to optimize SVM. The RBF kernel function is selected for the next step of parameter optimization.

A. GENETIC ALGORITHM-SUPPORT VECTORS MACHINES

GA is based on the natural evolutionary rules proposed by Darwin. The genetic algorithm has fast search capabilities and is easy to combine with other algorithms. It has excellent performance in many optimization problems [38].

In this paper, the accuracy of cross-validation is used as the fitness of GA. The higher the accuracy of c and g values under cross-validation is, the better the fitness is. The c and g with the highest classification accuracy are the global optimal parameters [39]. The population number is regarded as 20, and the termination algebra is given as 200. The probability of crossover is set as 0.9. The parameter c is selected from 0 to 100. The parameter g is selected from 0 to 1000. The flow chart is shown in Fig. 9.

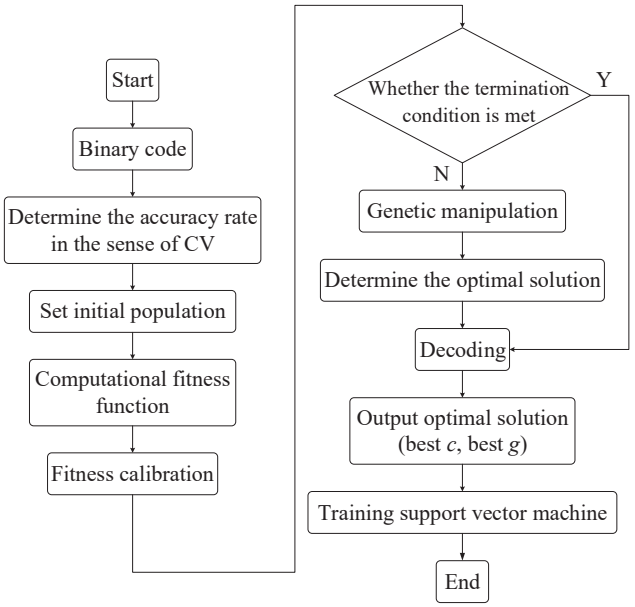
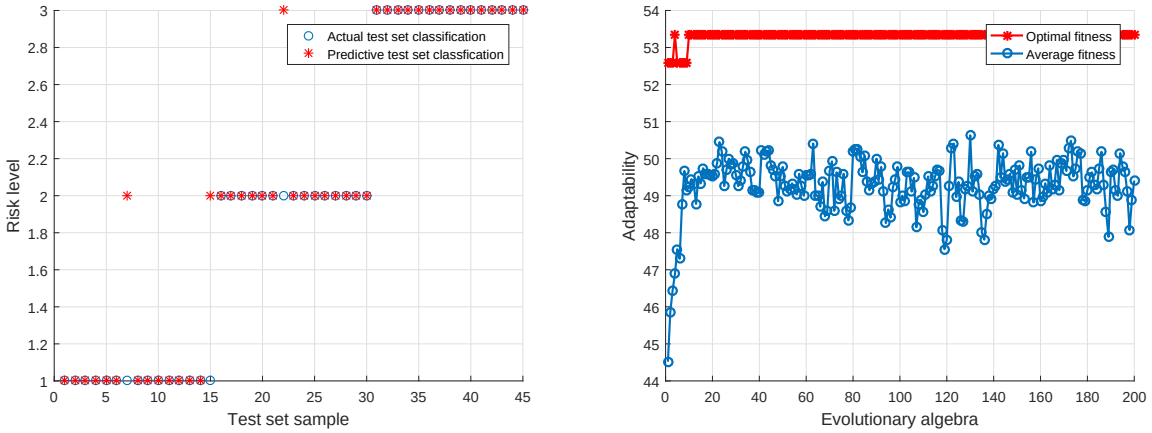


FIGURE 9. Flow chart of the GA-SVM.

Fig. 10(a) and Fig. 10(b) show the classification results and the fitness curve, respectively. After SVM is improved by GA, the best c is 10.029, the best g is 8.6162. For the prediction results of 45 samples in the test set, 13 cases are predicted correctly for the risk level 1, 14 cases are predicted correctly for the risk level 2, and 15 cases are predicted correctly for the risk level 3. The recognition accuracy rate rises to 93.33%.

In some related researches, SVM is used to predict breast cancer [40], chronic kidney disease [41], and freshwater disease [42] etc. GA is used to optimize neural networks [43] and traveling salesman problems [44]. In this paper, SVM is used to predict survival risks for patients with esophageal cancer. And GA is used to optimize SVM. The global optimization capabilities of GA is used to optimize the parameters of SVM, which improves the performance of SVM.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3007785, IEEE Access



(a) Prediction results based on GA-SVM.

(b) Fitness curve of GA optimization parameters.

FIGURE 10. GA-SVM.

B. PARTICLE SWARM OPTIMIZATION-SUPPORT VECTORS MACHINES

The basic principle of PSO is to simulate the behavior of the bird flock, and to find the best solution to the problem through information interaction within the group. The advantages of PSO includes simple operation, fast convergence, and global optimization. It has been widely used in many fields such as function optimization and image processing [45]. The position and velocity of each particle are updated as follows:

$$V_i^{k+1} = \omega V_i^k + c_1 r_1 \left(P_i - Z_i^k\right) + c_2 r_2 \left(P_g - Z_i^k\right) \quad (15)$$

$$Z_i^{k+1} = Z_i^k + V_i^{k+1} \quad (16)$$

where ω represtents the inertia factor, c_1 and c_2 are the acceleration constants. V_i and Z_i are the velocity vector and the position vector of the ith particle. P_i represtents the best neighborhood position. P_g is the best individual historical position. r_1 and r_2 represents the random number of the interval $[0, 1]$. k represents the number of iterations.

The population scale is set as 20, the learning factor c_1 is given as 2, the learning factor c_2 is selected as 2.5, the search range of penalty factor c is given as 0.1 to 100, the search range of kernel parameter g is selected from 0.01 to 1000, and the maximum number of iterations is regarded as 200. The flow chart is shown in Fig. 11.

Fig. 12(a) and Fig. 12(b) show the classification results and the fitness curve, respectively. After the parameters are improved by PSO, the best c is 25.3269, and the best g is 17.7279. Among the 45 test samples of three risk levels, 43 cases are correctly classified and predicted. The accuracy rate reaches 95.56%.

In the researches of some scholars, PSO is used to optimize the best drug use method [46], optimize detection of brain tumors [47] and analysis of epidemic models [48] etc. In this paper, a combination of PSO and SVM is realized to predict the survival risk of patients with esophageal cancer. PSO is used to optimize the parameters of SVM. As a result,

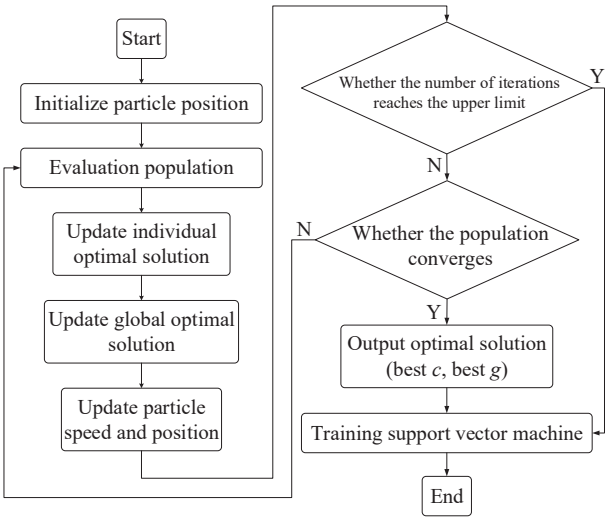


FIGURE 11. Flow chart of the PSO-SVM.

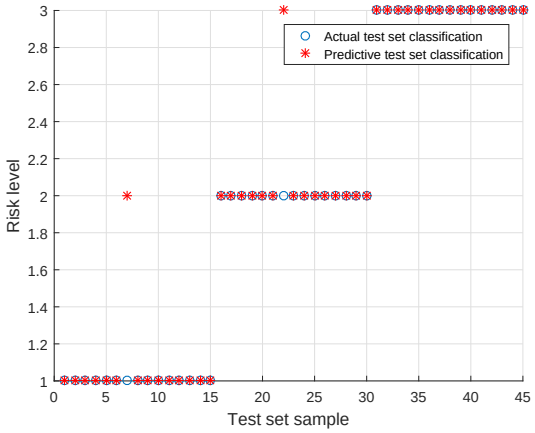
the classification performance of SVM is improved and the survival risk level prediction for esophageal cancer patients is improved.

C. ARTIFICIAL BEE COLONY-SUPPORT VECTORS MACHINES

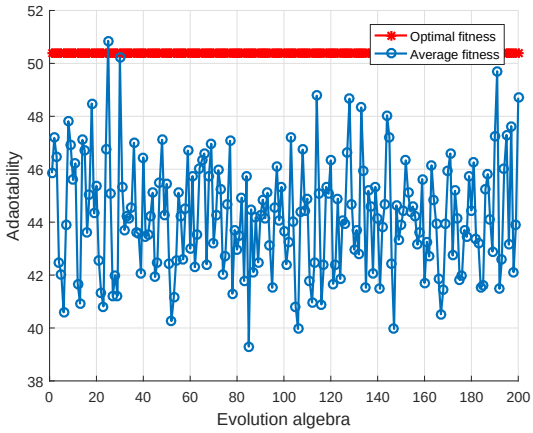
The artificial bee colony algorithm is an optimized method to simulate the behavior of bees, which can solve multivariate function optimization problems. It does not require specific information about the problem and has a faster convergence rate [49]. In this article, the ABC algorithm is used to optimize the main parameters of SVM, including penalty factor c and kernel function parameter g.

In this algorithm, the initial bee colony size is regarded as 20. The number of updates is limited to 100. If the honey source is not updated more than 100 times, the honey source is abandoned. The maximum number of iterations is set as 10

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3007785, IEEE Access



(a) Prediction results based on PSO-SVM.



(b) Fitness curve of PSO optimization parameters.

FIGURE 12. PSO-SVM.

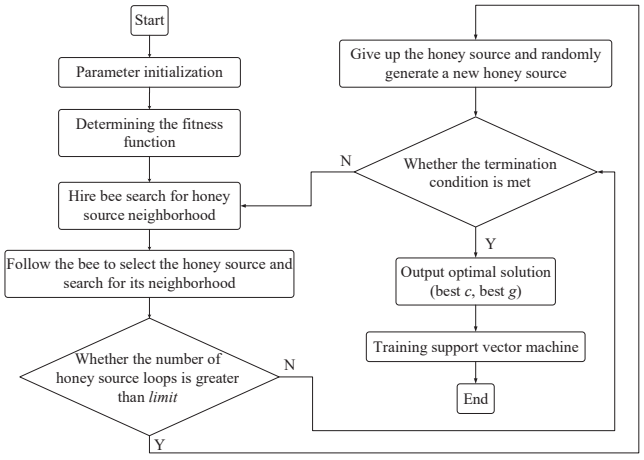


FIGURE 13. Flow chart of the ABC-SVM.

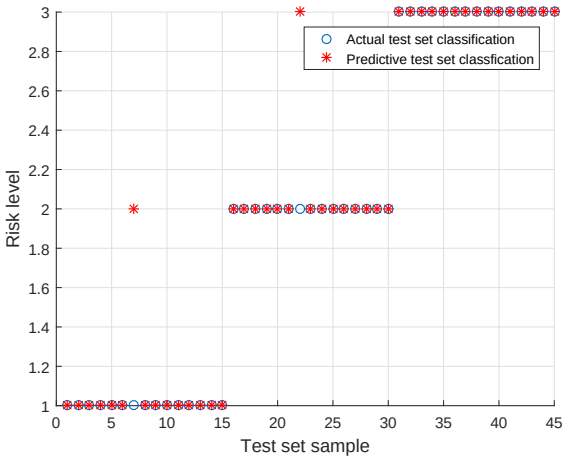


FIGURE 14. Prediction results based on ABC-SVM.

times. The number of parameters optimized is 2, the range of parameters is selected from 0.01 to 100, and the algorithm is repeated twice to check the robustness of the ABC-SVM. The flow chart is shown in Fig. 13.

The classification results are shown in Fig. 14. After the parameters optimization of the ABC, the best c is 10.6635, and the best g is 11.9966. Among the 45 test samples of three risk levels, 43 cases are correctly classified and predicted, and the accuracy rate reaches 95.56%.

In some related studies, ABC is used to optimize the text feature space [50], identify diseases in grape leaves [51] and optimize artificial neural networks [52]. In this paper, the combination of ABC and SVM is to predict the survival risk of esophageal cancer patients. ABC is used to optimize the parameters of SVM, and the performance of SVM is improved.

D. RESULTS ANALYSIS AND RESULTS DISCUSSION

In our study, the survival risk of esophageal cancer patients is well predicted. Firstly, nine blood indicators that are significantly related to the survival of patients with esophageal cancer are found. The SOM clustering algorithm is convenient and effective, and can cluster multiple blood indicators that are significantly related to the survival of esophageal cancer patients. Secondly, the patient's survival risk is divided into three risk levels. ROC not only has good two-classification performance, but also is insensitive to category imbalance. Based on the ROC method, the critical threshold for the survival of patients with esophageal cancer is calculated. Finally, the survival risk of esophageal cancer patients is predicted. In the study, different kernel functions of SVM are used for experiments and comparison. Three different optimization algorithms are used to optimize the parameters of SVM, and the performance of SVM is improved.

When the SVM has no parameter optimization, the prediction effect of RBF-SVM is better than Liner-SVM, Poly-SVM, and Sigmoid-SVM. When the parameters are optimized, the prediction accuracy of the SVM is improved. The

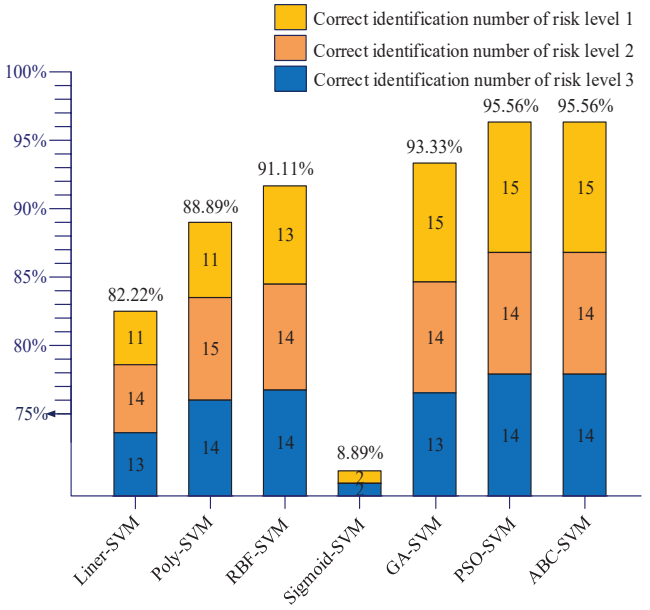


FIGURE 15. Prediction results of SVM ensembles.

optimization effect of ABC is better than that of GA and PSO. Fig. 15 shows the comparison of prediction results for Liner-SVM, Poly-SVM, RBF-SVM, Sigmoid-SVM, GA-SVM, PSO-SVM, and ABC-SVM.

Table 5 expresses the prediction results of survival risk levels of esophageal cancer based on different optimization algorithms. It is clear that the SVM algorithm based on parameter optimization is better than the unoptimized algorithm. The correct prediction rates of GA-SVM achieve 93.33%, while the correct prediction rates of PSO-SVM and ABC-SVM are all 95.56%. The parameter optimization effects of ABC and PSO are better than the GA, and ABC-SVM runs faster than other optimization algorithms.

VI. CONCLUSION

In order to predict and evaluate the survival risk levels of esophageal cancer accurately and efficiently, a method based on SOM clustering and SVM ensembles is proposed in this article. The SOM neural network, ROC curve, SVM, GA-SVM, PSO-SVM, and ABC-SVM are used in this method. The aim is to find more effective and accurate multiple blood indexes related to the survival of patients with esophageal cancer, and predictive classification of risk levels. All 180 esophageal cancer samples are investigated, nine blood indexes are extracted by SOM clustering, and three risk levels for survival are divided by ROC analyzing. Four kernel functions are used by SVM. RBF-SVM works better than Liner-SVM, Poly-SVM, and Sigmoid-SVM. The prediction accuracy of esophageal cancer is improved through three optimization algorithms. Among the algorithms studied in this paper, ABC-SVM has the best prediction rate and the shortest running time. It is concluded that ABC-SVM works better than that of GA-SVM and PSO-SVM in our study.

VII. ACKNOWLEDGMENT

This work was supported in part by the National Key Research and Development Program of China for International S and T Cooperation Projects under Grant 2017YFE0103900, in part by the Joint Funds of the National Natural Science Foundation of China under Grant U1804262, in part by the State Key Program of National Natural Science of China under Grant 61632002, in part by the Foundation of Young Key Teachers from University of Henan Province under Grant 2018GGJS092, in part by the Youth Talent Lifting Project of Henan Province under Grant 2018HYTP016, in part by the Henan Province University Science and Technology Innovation Talent Support Plan under Grant 20HASTIT027, in part by the Zhongyuan Thousand Talents Program under Grant 204200510003, and in part by the Open Fund of State Key Laboratory of Esophageal Cancer Prevention and Treatment under Grant K2020-0010 and Grant K2020-0011.

REFERENCES

[1] V. McCormack, D. Menya, M. Munishi, C. Dzamalala, N. Gasmelseed, M. Leon Roux, M. Assefa, O. Osano, M. Watts, A. Mwasamwaja et al., "Informing etiologic research priorities for squamous cell esophageal cancer in africa: a review of setting-specific exposures to known and putative risk factors," International Journal of Cancer, vol. 140, no. 2, pp. 259–271, 2017.

[2] I. Domingues, I. L. Sampaio, H. Duarte, J. A. Santos, and P. H. Abreu, "Computer vision in esophageal cancer: a literature review," IEEE Access, vol. 7, pp. 103 080–103 094, 2019.

[3] P. K. Varshney, H. Kumar, J. Kaur, and I. Gera, "Breast cancer risk prediction," IITM Journal of Management and IT, vol. 10, no. 1, pp. 37–45, 2019.

[4] P. Wang, Q. Song, Y. Li, S. Lv, J. Wang, L. Li, and H. Zhang, "Cross-task extreme learning machine for breast cancer image classification with deep convolutional features," Biomedical Signal Processing and Control, vol. 57, p. 101789, 2020.

[5] M. M. Saritas and A. Yasar, "Performance analysis of ann and naive bayes classification algorithm for data classification," International Journal of Intelligent Systems and Applications in Engineering, vol. 7, no. 2, pp. 88–91, 2019.

[6] I. M. Nasser and S. S. Abu-Naser, "Lung cancer detection using artificial neural network," International Journal of Engineering and Information Systems (IJEAIS), vol. 3, no. 3, pp. 17–23, 2019.

[7] M. Huang, C. Chen, W. Lin, S. Ke, and C. Tsai, "SVM and SVM ensembles in breast cancer prediction," PloS One, vol. 12, no. 1, p. e0161501, 2017.

[8] Y. Xiao, J. Wu, Z. Lin, and X. Zhao, "A deep learning-based multi-model ensemble method for cancer prediction," Computer Methods and Programs in Biomedicine, vol. 153, pp. 1–9, 2018.

[9] C. H. Jin, G. Pok, Y. Lee, H. W. Park, K. D. Kim, U. Yun, and K. H. Ryu, "A SOM clustering pattern sequence-based next symbol prediction method for day-ahead direct electricity load and price forecasting," Energy Conversion and Management, vol. 90, pp. 84–92, 2015.

[10] Z. Zhang, E. Lin, H. Zhuang, L. Xie, X. Feng, J. Liu, and Y. Yu, "Construction of a novel gene-based model for prognosis prediction of clear cell renal cell carcinoma," Cancer Cell International, vol. 20, no. 1, pp. 1–18, 2020.

[11] E. M. Ruiz, T. Niu, M. Zerfaoui, M. Kunnimalaiyaan, P. L. Friedlander, A. B. Abdel-Mageed, and E. Kandil, "A novel gene panel for prediction of lymph-node metastasis and recurrence in patients with thyroid cancer," Surgery, vol. 167, no. 1, pp. 73–79, 2020.

[12] T. Lan, Z. Xiao, H. Luo, K. Su, O. Yang, C. Zhan, and Y. Lu, "Bioinformatics analysis of esophageal cancer unveils an integrated mrna-lncrna signature for predicting prognosis," Oncology Letters, vol. 19, no. 2, pp. 1434–1442, 2020.

[13] F. Ni, Z. Lin, X. Fan, K. Shi, J. Ao, X. Wang, and R. Chen, "A novel genomic-clinicopathologic nomogram to improve prognosis prediction of hepatocellular carcinoma," Clinica Chimica Acta, 2020.

TABLE 5. Comparison of prediction results with parameter optimization.

Classification method	Risk levels	Correctly identify number	Recognition accuracy rate (%)	Best c	Best g	All sample number	All correctly identify number	Overall recognition rate (%)	Running times (S)
SVM(without optimization)	level 1	14	93.33	2	2	45	41	91.11	0.56
	level 2	14	93.33						
	level 3	13	86.67						
GA-SVM	level 1	13	86.67	10.029	8.6162	45	42	93.33	36.47
	level 2	14	93.33						
	level 3	15	100.00						
PSO-SVM	level 1	14	93.33	25.3269	17.7279	45	43	95.56	29.87
	level 2	14	93.33						
	level 3	15	100.00						
ABC-SVM	level 1	14	93.33	10.6635	11.9966	45	**43**	**95.56**	**1.38**
	level 2	14	93.33						
	level 3	15	100.00						

[14] C. H. Jin, G. Pok, Y. Lee, H.-W. Park, K. D. Kim, U. Yun, and K. H. Ryu, "A SOM clustering pattern sequence-based next symbol prediction method for day-ahead direct electricity load and price forecasting," Energy Conversion and Management, vol. 90, pp. 84–92, 2015.

[15] S. Delgado, C. Higuera, J. Calleespinosa, F. Morn, and F. Montero, "A SOM prototype-based cluster analysis methodology," Expert Systems with Applications, vol. 88, pp. 14–28, 2017.

[16] K. Matsuo, S. Purushotham, B. Jiang, R. S. Mandelbaum, T. Takiuchi, Y. Liu, and L. D. Roman, "Survival outcome prediction in cervical cancer: Cox models vs deep-learning model," American Journal of Obstetrics and Gynecology, vol. 220, no. 4, pp. 381–e1, 2019.

[17] J. Sun, X. Zhao, J. Fang, and Y. Wang, "Autonomous memristor chaotic systems of infinite chaotic attractors and circuitry realization," Nonlinear Dynamics, vol. 94, no. 4, pp. 2879–2887, 2018.

[18] C. Liang and J. Cai, "Bayesian variable selection and estimation in joint confirmatory factor analysis–cox model," Statistics and Its Interface, vol. 13, no. 1, pp. 49–63, 2020.

[19] B. Pache, M. Hübner, J. Solà, D. Hahnloser, N. Demartines, and F. Grass, "Receiver operating characteristic analysis to determine optimal fluid management during open colorectal surgery," Colorectal Disease, vol. 21, no. 2, pp. 234–240, 2019.

[20] Y. Wang, Z. Li, and J. Sun, "Three-variable chaotic oscillatory system based on dna strand displacement and its coupling combination synchronization," IEEE Transactions on NanoBioscience, vol. 19, no. 3, pp. 434–445, 2020.

[21] J. Sun, G. Han, Z. Zeng, and Y. Wang, "Memristor-based neural network circuit of full-function pavlov associative memory with time delay and variable learning rate," IEEE Transactions on Cybernetics, 2019, doi: 10.1109/TCYB.2019.2951520.

[22] R. Duarte, A. Stainthorpe, J. Greenhalgh, M. Richardson, S. Nevitt, J. Mahon, E. Kotas, A. Boland, H. Thom, T. Marshall et al., "Forest plots and summary receiver operating characteristic plots," in Lead-I ECG for Detecting Atrial Fibrillation in Patients with an Irregular Pulse using Single Time Point Testing: a Systematic Review and Economic Evaluation. NIHR Journals Library, 2020.

[23] P. Wen, S. Chen, J. Wang, and W. Che, "Receiver operating characteristics (roc) analysis for decreased disease risk and elevated treatment response to pegylated-interferon in chronic hepatitis b patients," Future Generation Computer Systems, vol. 98, pp. 372–376, 2019.

[24] T. Tada, T. Kumada, H. Toyoda, K. Tsuji, A. Hiraoka, K. Michitaka, A. Deguchi, T. Ishikawa, M. Imai, H. Ochi et al., "Impact of albumin–bilirubin grade on survival in patients with hepatocellular carcinoma who received sorafenib: An analysis using time-dependent receiver operating characteristic," Journal of Gastroenterology and Hepatology, vol. 34, no. 6, pp. 1066–1073, 2019.

[25] Y. Evcimen, I. U. Onur, H. Cengiz, and F. U. Yigit, "Optical coherence tomography findings in pre-eclampsia: a preliminary receiver operating characteristic analysis on choroidal thickness for disease severity," Current Eye Research, vol. 44, no. 8, pp. 916–920, 2019.

[26] E. J. Jang, B. Nandram, Y. Ko, and D. H. Kim, "Small area estimation of receiver operating characteristic curves for ordinal data under stochastic ordering," Statistics in Medicine, 2020.

[27] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu, "Applications of support vector machine (SVM) learning in cancer genomics," Cancer Genomics-Proteomics, vol. 15, no. 1, pp. 41–51, 2018.

[28] K. Hou, G. Shao, H. Wang, L. Zheng, Q. Zhang, S. Wu, and W. Hu, "Research on practical power system stability analysis algorithm based on modified SVM," Protection and Control of Modern Power Systems, vol. 3, no. 1, p. 11, 2018.

[29] M. M. Elsaadawi and A. Y. Hatata, "A novel protection scheme for synchronous generator stator windings based on SVM," Protection and Control of Modern Power Systems, vol. 2, no. 1, pp. 1–12, 2017.

[30] J. Zhou, L. Li, L. Wang, X. Li, H. Xing, and L. Cheng, "Establishment of a SVM classifier to predict recurrence of ovarian cancer," Molecular Medicine Reports, vol. 18, no. 4, pp. 3589–3598, 2018.

[31] S. Bandyopadhyay, S. Saha, U. Maulik, and K. Deb, "A simulated annealing-based multiobjective optimization algorithm: Amosa," IEEE Transactions on Evolutionary Computation, vol. 12, no. 3, pp. 269–283, 2008.

[32] T. Zhu, S. Chen, W. Zhu, and Y. Wang, "Optimization of sound absorption property for polyurethane foam using adaptive simulated annealing algorithm," Journal of Applied Polymer Science, vol. 135, no. 26, p. 46426, 2018.

[33] Y. Cai, "Artificial fish school algorithm applied in a combinatorial optimization problem," International Journal of Intelligent Systems and Applications, vol. 2, no. 1, pp. 37–43, 2010.

[34] R. Jovanovic, M. Tuba, and S. Vos, "An efficient ant colony optimization algorithm for the blocks relocation problem," European Journal of Operational Research, vol. 274, no. 1, pp. 78–90, 2019.

[35] S. Mirjalili and A. Lewis, "The whale optimization algorithm," Advances in Engineering Software, vol. 95, no. 95, pp. 51–67, 2016.

[36] I. Aljarah, H. Faris, and S. Mirjalili, "Optimizing connection weights in neural networks using the whale optimization algorithm," Soft Computing, vol. 22, no. 1, pp. 1–15, 2018.

[37] A. Darwish, "Bio-inspired computing: Algorithms review, deep analysis, and the scope of applications," Future Computing and Informatics Journal, vol. 3, no. 2, pp. 231–246, 2018.

[38] C. Sukawattanavijit, J. Chen, and H. Zhang, "GA-SVM algorithm for improving land-cover classification using SAR and optical remote sensing data," IEEE Geoscience and Remote Sensing Letters, vol. 14, no. 3, pp. 284–288, 2017.

[39] Z. Tao, L. Huiling, W. Wenwen, and Y. Xia, "GA-SVM based feature selection and parameter optimization in hospitalization expense modeling," Applied Soft Computing, vol. 75, pp. 323–332, 2019.

[40] G. Verma and H. Verma, "Predicting breast cancer using linear kernel support vector machine," Available at SSRN 3350254, 2019.

[41] N. A. Almansour, H. F. Syed, N. R. Khayat, R. K. Altheeb, R. E. Juri, J. Alhiyafi, S. Alrashed, and S. O. Olatunji, "Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study," Computers in Biology and Medicine, vol. 109, pp. 101–111, 2019.

[42] H. Hassani, E. S. Silva, M. Combe, D. Andreou, M. Ghodsi, M. R. Yeganegi, and R. E. Gozlan, "A support vector machine based approach for predicting the risk of freshwater disease emergence in england," Stats, vol. 2, no. 1, pp. 89–103, 2019.

[43] M. S. H. Kalathingal, S. Basak, and J. Mitra, "Artificial neural network modeling and genetic algorithm optimization of process parameters in fluidized bed drying of green tea leaves," Journal of Food Process Engineering, p. e13128, 2020.

[44] T. George and T. Amudha, "Genetic algorithm based multi-objective optimization framework to solve traveling salesman problem," in Advances in Computing and Intelligent Systems. Springer, 2020, pp. 141–151.

[45] C.-L. Huang and J.-F. Dun, "A distributed PSO–SVM hybrid system with feature selection and parameter optimization," Applied Soft Computing, vol. 8, no. 4, pp. 1381–1391, 2008.

[46] O. Shindi, J. Kanesan, G. Kendall, and A. Ramanathan, "The combined effect of optimal control and swarm intelligence on optimization of cancer chemotherapy," Computer Methods and Programs in Biomedicine, p. 105327, 2020.
[47] M. Sharif, J. Amin, M. Raza, M. Yasmin, and S. C. Satapathy, "An integrated design of particle swarm optimization (PSO) with fusion of features for detection of brain tumor," Pattern Recognition Letters, vol. 129, pp. 150–157, 2020.
[48] M. Mahmoodabadi, "Epidemic model analyzed via particle swarm optimization based homotopy perturbation method," Informatics in Medicine Unlocked, p. 100293, 2020.
[49] H. M. Alshamlan, G. H. Badr, and Y. A. Alohali, "ABC-SVM: artificial bee colony and SVM method for microarray gene selection and multi class cancer classification," Int. J. Mach. Learn. Comput, vol. 6, no. 3, p. 184, 2016.
[50] P. Grover and S. Chawla, "Text feature space optimization using artificial bee colony," in Soft Computing for Problem Solving. Springer, 2020, pp. 691–703.
[51] A. D. Andrushia and A. T. Patricia, "Artificial bee colony optimization (ABC) for grape leaves disease detection," Evolving Systems, pp. 1–13, 2019.
[52] J. Watada, A. Roy, B. Wang, S. C. Tan, and B. Xu, "An artificial bee colony based double layered neural network approach for solving quadratic bi-level programming problems," IEEE Access, 2020.

JUNWEI SUN received the Ph.D. degree from Huazhong University of Science and Technology, Wuhan, China, in 2014. Since 2014, he has been with the School of Electrical and Information Engineering, Zhengzhou University of Light Industry. He was an Assistant Professor, became an Associate Professor in 2018. He has published over 50 SCI journal papers. His research interests include data processing and analysis, complex networks, and memristor-based neural network.

YULI YANG received the B.Eng. degree from Henan Agricultural University, Zhengzhou China, in 2014 and 2018. Since 2018, he has been working toward the MA.Eng. degree in control engineering at the School of Electrical and Information Engineering, Zhengzhou University of Light Industry, Zhengzhou, China. His research interests include data processing and computer modeling.

YANFENG WANG received the M.S. and Ph.D. degrees from Huazhong University of Science and Technology, Wuhan, China, in 2004 and 2007, respectively. He is currently a Professor with the School of Electrical and Information Engineering, Zhengzhou University of Light Industry, Zhengzhou, China. He has published over 80 SCI journal papers in the areas of dynamic modelling, data driven modeling, pattern recognition, and control and synchronization control. His research interests include data analysis and computer modeling.

LIDONG WANG received the Ph.D. degrees from Zhejiang University, Hangzhou, China, in 2004. He is currently a Professor with the Zhengzhou University, the director for State key laboratory for esophageal cancer prevention and treatment & Henan key laboratory for esophageal cancer research of the first affiliated hospital, Zhengzhou University. He has published 570 papers as esophageal cancer research, and 23 papers of them were classified as JCR 1.

XIN SONG received her M.S. degree from Zhengzhou University, Zhengzhou, China, in 2008. She is currently a Ph.D. student for esophageal cancer research in Zhengzhou University. She has published 16 SCI journal papers in the areas of esophageal cancer research.

XUEKE ZHAO received the Ph.D. degrees from Zhengzhou University in 2017. He is currently an assistant research fellow with State Key Laboratroy of Esophgeal Cancer Preventive & Treatment and Henan Key Laboratroy for Esophgeal Cancer Research of The First Affiliated Hospital, Zhengzhou University, Henan Province, PR China. He has published more than 40 papers in the areas of pathogenesis and prevention of esophageal cancer. His interest focuses on the molecular typing of esophageal cancer and the definition of high-risk groups for esophageal cancer.