

Research Article

Auxiliary Medical Decision System for Prostate Cancer Based on Ensemble Method

Jia Wu ^{1,2}, Qinghe Zhuang,^{1,2} and Yanlin Tan ^{2,3}

¹School of Computer Science and Engineering, Central South University, Changsha 410083, China

²“Mobile Health” Ministry of Education-China Mobile Joint Laboratory, Changsha 410083, China

³PET-CT Center, The Second Xiangya Hospital of Central South University, Changsha 410083, China

Correspondence should be addressed to Jia Wu; jiawu5110@163.com and Yanlin Tan; tanyanlin@csu.edu.cn

Received 3 March 2020; Accepted 27 April 2020; Published 18 May 2020

Guest Editor: Plácido R. Pinheiro

Copyright © 2020 Jia Wu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Prostate cancer (PCa) is one of the main diseases that endanger men’s health worldwide. In developing countries, due to the large number of patients and the lack of medical resources, there is a big conflict between doctors and patients. To solve this problem, an auxiliary medical decision system for prostate cancer was constructed. The system used six relevant tumor markers as the input features and employed classical machine learning models (support vector machine and artificial neural network). Stacking method aimed at different ensemble models together was used for the reduction of overfitting. 1,933,535 patient information items had been collected from three first-class hospitals in the past five years to train the model. The result showed that the auxiliary medical system could make use of massive data. Its performance is continuously improved as the amount of data increases. Based on the system and collected data, statistics on the incidence of prostate cancer in the past five years were carried out. In the end, influence of diet habit and genetic inheritance for prostate cancer was analyzed. Results revealed the increasing prevalence of PCa and great negative impact caused by high-fat diet and genetic inheritance.

1. Introduction

In 2018, morbidity and mortality of PCa accounted for 13.5% and 6.7%, respectively, in male patients. In 185 countries around the world, PCa has the highest morbidity in 105 countries and the highest mortality in 46 countries [1]. Undoubtedly, PCa has become one of the main threats to men’s health worldwide. Countries with high HDI (human development index) usually have high morbidity (68.0/100,000). Although countries with low HDI have relatively low morbidity (14.5/100,000), there is an obvious trend of growth and it increases fast [2]. Take China for example; in 1998, the rough morbidity was 3.25/100,000; however, in 2012, it increased to 8.14/100,000. By 2013, it has become 8.58/100,000 [3, 4]. The high morbidity in developed countries and the increasing incidence rate in developing countries have led to a huge number of prostate cancer patients worldwide.

In developing countries that lack medical resources, many patients cannot receive timely and effective diagnosis

and therapy, which will aggregate the conflict between doctors and patients. In China, there are only 2.59 practitioners for every 1000 people [5]. In Beijing, a small number of high-level hospital medical staff members have to provide medical services to more than 20 million people in Beijing, and many cancer patients in other regions also come here for the extraordinary medical treatment. Medical staff members have been overloaded for a long time [6–9]. This will inevitably have an impact on the accuracy of the diagnosis, thus leading to serious consequences.

Other developing countries may face similar problems:

- (i) Due to the large number of patients and less medical resources, it is difficult for patients to get timely and effective diagnosis and treatment
- (ii) The long-term workload of doctors reduces the efficiency and accuracy of diagnosis
- (iii) Many hospitals have poor medical equipment, which further increases the probability of misdiagnosis

- (iv) The per capita income in developing countries is pretty low, and most people cannot afford expensive but accurate diagnostic methods like PET-CT

These problems can be alleviated by building an auxiliary medical decision system. By analyzing a large number of data, the auxiliary medical decision system can learn a diagnostic model. When a new patient comes, it will provide doctors with suggestions relevant to diagnosis or treatment based on the learned model [10, 11]. Combing suggestions from the system and their own knowledge, doctors will give the final conclusion of diagnosis and treatment method. The auxiliary medical decision system can relieve the doctor's burden to some extent, thus alleviating the conflict between doctors and patients [12, 13]. In this work, we constructed an auxiliary medical decision system which can determine whether a patient has prostate cancer, judge the clinical stage, recommend treatment options, and evaluate the effectiveness of treatment options. Given the low income in developing countries, six tumor markers are selected with relatively low testing price and high relevance to PCa as the input features. Classical machine learning techniques and ensemble method are adopted to extract the knowledge inside data and improve performance.

The main contributions and innovations of this research include the following:

- (i) Appropriate features are selected for the construction of medical systems according to national conditions of developing countries
- (ii) The constructed auxiliary system can give treatment plan and evaluate its effectiveness
- (iii) The use of constructed ensemble method by a secondary learner improved the accuracy
- (iv) The system is trained based on a large amount of patient information from three high-level hospitals in China, and some factors affecting PCa via the constructed system are analyzed

The main structure of this article is as follows: the first part describes the background and contribution of the research, the second part introduces related research, the third part is a detailed description to the construction of the model, the fourth part is about the training process and analysis to experiments' results, and the fifth part is the conclusion.

2. Related Works

Medical diagnosis of cancer is usually a gradual transition which starts from simple, cheap, and harmless but with low-accuracy methods and ends with expensive and accurate methods. Compared with commonly used CT, MRI, PET-CT, and other methods, detection of tumor markers is a relatively basic and cheap method in the diagnosis of cancer, which makes constructing the auxiliary medical decision system with tumor markers suitable for developing countries with low capital medical expenditure, low medical level,

and poor medical facilities. Diagnosing cancer with a single tumor marker usually does not have good sensitivity and specificity [14]. Therefore, many related research studies [14–17] combine different tumor markers or biomarkers to predict some diseases and have acquired good results. Specifically, literature [15] detected serum levels of 17 tumor markers for 145 patients with pancreatic cancer and selected 9 tumor markers by backward elimination selection, scatter plots, and relative operating characteristic analysis. Based on these features, the pancreatic cancer determination system CAMPAS-P was established. The final results showed that the CAMPAS-P system was able to accurately distinguish malignant pancreatic cancer from benign pancreatobiliary disease but performed bad on the diagnosis of the unusual histologic type of pancreatic tumors and various digestive organ malignancies. Literature [17] used serum microRNA biomarkers to predict nonalcoholic fatty liver disease (NAFLD). Among all the selected biomarkers, nine of them were associated with NAFLD severity, and some of them appeared specific to NAFLD. These biomarkers showed good classification performance for nonalcoholic steatohepatitis (NASH). Literature [14] combined growth-related tumor markers and associated tumor markers for the diagnosis of cancer and acquired 80–90% sensitivity, 84–85% specificity, and 83–88% accuracy.

As one of the classic machine learning algorithms, the support vector machine (SVM) [18] shows good performance in many classification problems before the revival of deep learning. It is also widely used in medical diagnosis [19–23]. Literature [19] extracted features from mammograms by Hough transform and classified mammograms by SVM. Its accuracy reached 94% while other machine learning methods like linear discriminant analysis just had 86% accuracy. Literature [20] employed various machine learning techniques for the prediction of breast cancer in Wisconsin Breast Cancer (original) datasets. After considering accuracy, sensitivity, specificity, and precision, SVM got the best results. Literature [21] proposed a classification fuzzy-rough set with the SVM model featured as CA-125 and other amino acids to detect early-stage ovarian cancer. It performed quick learning and had good classification performance.

Ensemble learning technology is also widely used in medical assistant diagnosis. Literature [24] selected demographic, physiological, and vital signs and laboratory tests as features and built different models, finding that the ensemble learning random forest is the most effective in mortality prediction in the early hours of an ICU patient admission. The proposed EMPICU-RF framework based on the ensemble learning random forest outperformed many standard scoring systems in terms of AUROC (area under the curve) and time. Literature [25] proposed a model that combines the physicians' knowledge in the form of a rule-based classifier and supervises learning algorithms to detect asthma control level. Literature [23] constructed two different ensemble models by confidence-weighted voting method and the boosting ensemble technique for the diagnosis of breast cancer. The proposed CWV-BANN-SVM model reached the accuracy of 100%.

TABLE 1: Commonly used kernel functions.

Kernel functions	Formula
Linear kernel	$\kappa(x_i, x_j) = x_i^T x_j$
Polynomial kernel	$\kappa(x_i, x_j) = (x_i^T x_j)^d$
Gauss kernel	$\kappa(x_i, x_j) = \exp\left(-\frac{\ x_i - x_j\ ^2}{2\sigma^2}\right)$
Sigmoid kernel	$\kappa(x_i, x_j) = \exp\left(-\frac{\ x_i - x_j\ }{\sigma}\right)$
Laplace kernel	$\kappa(x_i, x_j) = \tanh(\beta x_i^T x_j + \theta)$

3. Design of the Auxiliary Medical Decision-Making System

3.1. Requirements and Framework of the System. The auxiliary medical decision system is aimed at offering some help to doctors. Its functions cover diagnosing the patients, staging the cancer, recommending the treatment plan, and evaluating the treatment plan. Diagnosing patients is to tell if the tumor is malignant or benign. Staging the cancer is to determine the clinical stage (I, II, III, or IV) for those diagnosed with malignant PCa. The above two requirements can be satisfied by building a classification model using the machine learning method. On the other hand, in order to give a cancer treatment plan and evaluate its efficacy at the same time, the whole problem is considered a regression problem. The system will finally output a value evaluating the malignancy of PCa, abbreviated as EM value. The larger the value, the higher the malignancy. If the value does not decrease after executing a certain treatment plan, it means that the treatment plan is not effective and another treatment plan needs to be selected. Meanwhile, the auxiliary medical decision-making system needs to have good parallelism and be able to process multiple patients' simultaneous diagnosis requests. It is worth noting that after the medical system is invested, the amount of data obtained will gradually increase over time. The decision model will be retrained to further improve the generalization performance.

3.2. Design of the Decision Model

3.2.1. Introduction to the Support Vector Machine. The support vector machine can acquire the global optimal solution in high-dimensional problems; thus, it is widely used in many situations [26].

For a linear separable binary classification problem, assume that the input dataset $S = \{x_1, x_2, x_3, \dots, x_m\}$ and the output label $y = \{y_1, y_2, y_3, \dots, y_m\}$, where x_i is the input vector of the i th sample and $y_i \in \{-1, 1\}$ is the corresponding label of x_i . SVM aims to find a hyperplane $H : W^T x + b = 0$ that separates the positive and negative samples and meanwhile maximizes their distance to the hyperplane. The optimizing process can be expressed as follows:

$$\begin{cases} \min_{w,b} \frac{1}{2\|W\|^2}, \\ y_i(W^T x + b) \geq 1, \quad i = 1, 2, \dots, m. \end{cases} \quad (1)$$

In order to compute the solution efficiently, the Lagrange multiplier is introduced, and according to the Wolfe duality theory, it is changed into an equivalent dual problem:

$$\text{s.t. } \sum_{i=1}^m y_i \alpha_i = 0, \quad \alpha_i \geq 0, \quad i = 1, 2, \dots, m. \quad (2)$$

As for the linear inseparable problem, penalty parameter C and slack variable ξ_i are introduced, and the problem is expressed as follows:

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m y_i y_j \alpha_i \alpha_j x_i^T x_j - \sum_{j=1}^n \alpha_j \\ \text{s.t.} \quad & \sum_{j=1}^m y_j \alpha_j = 0 \quad 0 \leq \alpha_i \leq C, \quad i = 1, 2, \dots, m, \end{aligned} \quad (3)$$

where C represents the interval of two classes, and the final decision function is as follows:

$$y(x) = \text{sgn}(f(x)) = \text{sgn} \left\{ \left(\sum_{i=1}^m \alpha_i^* y_i (x_i^T \cdot x) \right) + b^* \right\}. \quad (4)$$

For the nonlinear problem, kernel function $\varphi(x_i)$ that maps the nonlinear problem in low dimension into a linear problem in high dimension is introduced. The kernel functions defined in the input sample space satisfy the Mercer condition and have the following expression: $K(x_i, x_j) = (\varphi(x_i), \varphi(x_j)) = \varphi(x_i)^T \varphi(x_j)$. Commonly used kernel functions are listed in Table 1.

A typical support vector machine is usually used to deal with binary classification problems. In this medical decision system, SVM is firstly used to classify benign (labeled 1) and malignant (labeled -1) tumors. In order to stage possible malignant tumors, a four-class (I, II, III, or IV) classification task is completed by one-to-one method, which means to train $\binom{4}{2}$ SVMs simultaneously and integrate the results of each SVM by majority voting. In the training process of one-to-one method, each SVM only needs the data in two classes which will have smaller training cost and keep the generalization performance at the same time.

The selection of the kernel function is one of the main factors that influence the performance of SVM. Commonly used kernel functions include linear function, polynomial function, sigmoid function, and radial basis function. Here, multiple kernel functions are used simultaneously to construct the SVM-based multiclassifiers in case of poor generalization performance due to wrong kernel function selection.

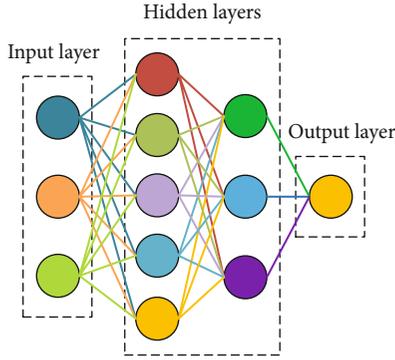


FIGURE 1: Schematic diagram of MLP.

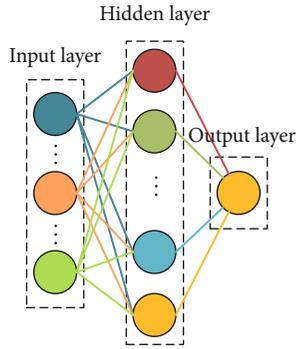


FIGURE 2: Schematic diagram of the RBF neural network.

3.2.2. Introduction to the Neural Network. The neural network (NN) is a model with strong fitting ability and is widely used by researchers in various disciplines. It is mainly composed of an input layer, hidden layers, and an output layer. According to the structure of the network, NN can be divided into multiple types: multilayer perceptron (MLP) neural network, radial basis function (RBF) neural network, adaptive resonance theory (ART) neural network, self-organizing map (SOM) neural network, etc. There are two main network models used in this medical decision system, MLP neural network and RBF neural network.

The MLP neural network contains one input layer, one or more hidden layers, and one output layer. Usually, every two adjacent layers are fully connected as shown in Figure 1.

The activation function of the MLP neural network is usually the rectified linear unit or the ReLU function which can be expressed as $f(x) = \max(0, x)$.

The RBF neural network is another widely used neural network [27]. It converges fast and has strong generalization ability. Unlike the MLP neural network, the RBF network contains only one hidden layer and uses radial basis function $f(x, c_i) = \exp(-\beta_i \|x - c_i\|^2)$ as the activation function, where c_i is the center of the i th unit in the hidden layer. The structure of the RBF neural network is shown in Figure 2.

3.2.3. Ensemble Learning. Ensemble learning is a method that integrates many base learners together to improve the overall learning ability. Commonly used ensemble learning methods include boosting, bagging, and stacking. The stacking method can generate all base learners in parallel

and uses a secondary learner to integrate the results of the base learners. This integration method has suitable training cost and strong generalization performance. In our medical decision-making system, in order to provide corresponding treatment plan recommendations and evaluate the efficacy of the plan after staging the cancer, the results of the classifier needs to be converted into a regression value. In this case, stacking is a good choice. To make ensemble method really work, base learners need to have some difference. This difference may come from different models, different input datasets, or different input features. Given that SVM and neural networks are less sensitive to the input dataset and our feature set is small, it is not suitable to use different input samples or different input features to improve generalization performance. However, there are many choices in the selection of the kernel function for SVM and structure for neural networks. Therefore, parameter perturbation is taken to enlarge the difference between base learners so that our decision model can be strengthened. Finally, the weights of base learners are learned through exponential linear regression (ELR) to obtain the evaluation of tumor malignancy or the EM value.

3.3. Detailed Description of the Medical Decision System. In the proposed medical decision system, six important tumor markers including prostate-specific antigen (PSA), prostate-specific membrane antigen (PSMA), total prostate-specific antigen (tPSA), red blood cell (RBC), hemoglobin (HB), and prostate acid phosphatase (PAP) are chosen for the diagnosis of PCa as the input features and SVM as the diagnosis model. Clinical stage determination of malignant PCa, treatment recommendation, and evaluation are completed by an ensemble model that combines SVM groups for four-class classification and neural networks with different structures. Figure 3 depicts the main flow of the auxiliary medical system.

First, relevant data from different hospital systems are collected. Then, six important tumor markers' levels are extracted from thousands of information items. After dropping samples with missing or abnormal value, an input vector $x = (x_{\text{PSA}}, x_{\text{PSMA}}, x_{\text{tPSA}}, x_{\text{RBC}}, x_{\text{HB}}, x_{\text{PAP}})$ is formed. Next, it will firstly use SVM to judge if the tumor is malignant. In clinical medicine, the increase in tumor marker level does not mean the development of a malignant tumor for sure. Many benign lesions or inflammations may also lead to an increase in tumor marker level, but the increase is not large. When the system determines that the tumor is benign, recommendations about the next examination and corresponding treatment will be listed.

If the tumor is judged to be malignant, the ensemble model will complete the stage division. The development of the malignant tumor will be divided into four stages: I, II, III, and IV; that is, to say, the system must complete a four-classification task. Since SVM is mainly used at binary classification problems, one-to-one strategy is taken and every six SVM models form a group of SVMs. The output of each group is a voting combination of six binary SVM classifiers in the group, which is represented by a four-dimensional one-hot vector. The differentiation of the SVM classifier is realized by choosing different kernel functions to improve the final performance. More explicitly, SVMs in the same

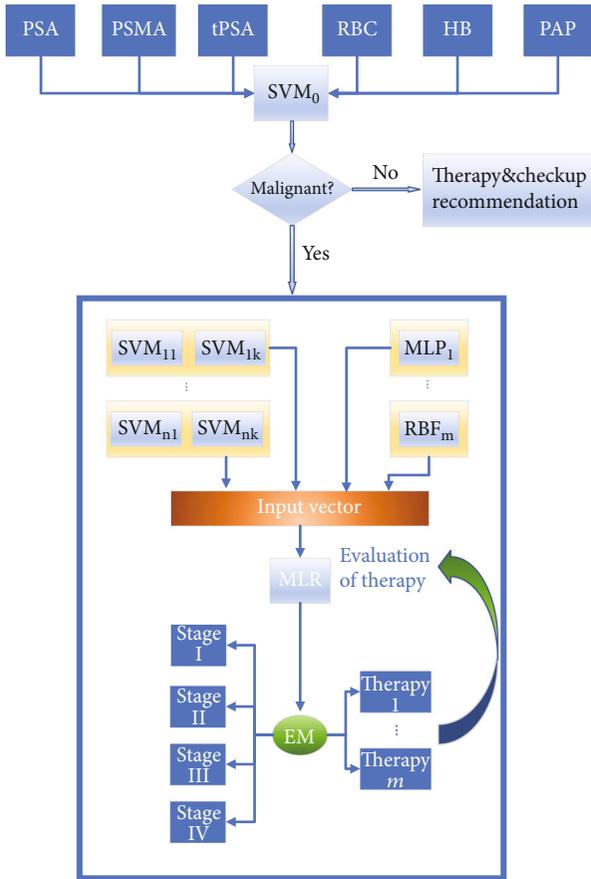


FIGURE 3: The main flow of the auxiliary medical decision system.

group use the same kernel function, and SVMs in different groups use different kernel functions. Three commonly used kernel functions: linear kernel, polynomial kernel, and Gaussian kernel are chosen to differentiate SVM groups.

While training, parameters in the kernel function and penalty parameters of each binary SVM are adjusted to reduce the generalization error below the threshold ϵ . In order to further reduce the risk, the widely used MLP neural network and RBF neural network are added into the system. Because 6 input features are selected and samples are classified into four classes, the input and output layers of the MLP and RBF networks are 6 units and 4 units, respectively. Three group MLP neural networks with different structures are selected: 6-9-7-4, 6-10-7-5-4, and 6-7-5-4 (these numbers follow the order of unit numbers in each layer). The ReLU function is used as the activation function in MLP neural networks. Similarly, three RBF networks with different structures are used. The hidden unit numbers in three networks are set as 10, 14, and 16, respectively. After clustering the samples by k -means algorithm, the center c_i of each hidden unit is determined. The activation function of the RBF neural networks is determined to be radial basis function. For the MLP and RBF networks, the hyperparameters are adjusted to reduce the generalization error below the threshold ϵ .

Finally, outputs of each SVM group and all MLP and RBF networks are connected into one vector, which will be the

input of the secondary learner. By observing the tumor marker level in the dataset, it is found that for benign tumors and patients in stage I, tumor marker levels are usually close to the normal range. But for patients in stage III and stage IV, the level of tumor markers deviates greatly from the normal range. Therefore, we assume that the growth of tumor markers in the development process of PCa conforms to the exponential law. This hypothesis is basically true in medicine. In the early stage, symptoms are very slight or not obvious. Tumors tend to be hard to find and grow slowly. However, in the middle and late stages, they grow savagely and spread throughout the body, making tumor marker levels really high.

Therefore, ELR is selected as the secondary learner to ensemble results of the SVM, MLP, and RBF models. Supervising output values 3, 4, 5, and 6 are added manually for input patient samples in stages I, II, III, and IV, respectively. What may be wired is that the output value of ELR is not set to start from 1. It is considered for the reason of improving the model's robustness to normal people and benign tumor cases. Finally, the evaluation value of PCa's malignancy (EM value) is output. Algorithm 1 shows the procedure that integrates the results of base learners by ELR.

The system determines the stage of malignant PCa according to the EM value and recommends the appropriate treatment method. Commonly used tumor treatment methods include chemotherapy, radiotherapy, excision, drug method, and hospital charge. After treatment methods are recommended by the system, doctors will decide to take it or abandon it or make modification based on it. What must be emphasized is that in the aspect of therapy recommendation, the system mainly gives a rough strategy to relieve doctors' pressure to some extent instead of replacing doctors completely. Concrete treatment is relevant to many factors which cannot be simply given by the system. If one patient has been treated for a while, the system will track the changes of his or her tumor marker levels, input the results of the tumor marker test into the system, and output the EM value to evaluate effect of the treatment. If the EM value changes a little, the treatment plan will be changed. If the EM value decreases greatly, this means it really works, so the original plan will be maintained. During this process, every effective therapy and corresponding EM value will be recorded in the database for further use.

4. Experiment

4.1. Dataset and Models' Training. We collected a large amount of data from three top-class hospitals in China: First Xiangya Hospital, Second Xiangya Hospital, and Third Xiangya Hospital. Relevant information about the data is shown in Table 2.

After screening and preprocessing the data, relevant records of the tumor markers (PSA, PSMA, tPSA, RBC, HB, and PAP) and diagnostic results (benign, stage I, stage II, stage III, and stage IV) are obtained.

Table 3 shows the normal range of six tumor markers related to PCa. Values of malignant patients' tumor marker are several times or even tens of times beyond the normal range.

```

Input:
Training set  $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}, y_i \in \{I, II, III, IV\}$ 
Primary classifier  $S = \{SVM_1, SVM_2, SVM_3, MLP_1, MLP_2, MLP_3, RBF_1, RBF_2, RBF_3\}$ 
Output:
Second learning algorithm  $H(x)$ :  $\ln(y_{EM}) = w^T x + b$ 
/*
 $x$  is the input vector.
 $w \in R^{36 \times 1}$  is the weight vector, and each element represents the corresponding weight in  $x$ ;  $w^T$  is the transpose of  $w$ .
 $b \in R^{36 \times 1}$  is the bias vector.
 $y_{EM}$  is the output EM value.
*/
Begin
 $D' = \emptyset$ 
for  $i$  in  $D$  do
  for  $t$  in  $S$  do
     $z_{it} = S_t(x_i)$ ; /*  $z_{it}$  is a four-dimensional vector. */
  end for
   $y'_i = \text{map}(y_i)$ ; /* map function converts the class label into a numerical value. */
   $D' = D' \cup ((z_{i1}, z_{i2}, \dots, z_{it}), y'_i)$ ;
end for
use  $D'$  to train  $H(x)$ ;
output  $H(x)$ ;
End

```

ALGORITHM 1

TABLE 2: Type and number of collected data.

Data type	Number
Patient information	1,933,535 items
Outpatient service	691,238 people
Doctors' device in outpatient	24,021,298 items
Be hospitalized	1,149,187 people
Diagnosis	1,089,327 items
Electronic medical records	4,855,619 items
Doctors' device in clinical	25,757,699 items
Inspection records	157,426 items
Medical laboratory records	8,725,586 items
Routine inspection records	22,358,881 items
Operation records	318,022 items
Drug records	120,546 items

Figure 4 shows the training process. The datasets are divided into two parts: training set and test set, accounting for 80% and 20%, respectively. Each of them is then divided into $S_{\text{malignant}}$ and S_{benign} . First of all, $S_{\text{malignant}}$ and S_{benign} are used and the appropriate kernel function and penalty parameter are searched to train SVM_0 until the test error is below ϵ . Second, malignant samples are divided into four parts $S_{\text{malignant}} = \{S_I, S_{II}, S_{III}, S_{IV}\}$ according to their clinical stages. SVM and neural networks are not sensitive to data. What is more, arbitrary division of data is likely to lead to the problem of imbalanced data which means two datasets do not have the same distribution. Hence, the whole training set is used to train all base learners instead of dividing it into several parts. Each binary SVM is trained separately. Majority

TABLE 3: Normal range of different tumor markers.

Types of tumor marker	Normal range
Prostate-specific antigen	0-4.0 ng/mL
Total prostate-specific antigen	4-20 $\mu\text{g/L}$
Hemoglobin	120-165 g/L
Red blood cell	12-15 g/100 mL
Prostate acid phosphatase	0-9 U/L
Prostate-specific membrane antigen	0-4 ng/mL

voting is used to ensemble the output results of SVMs in the same group. For neural networks, the malignant samples are directly marked as $(1, 0, 0, 0)$, $(0, 1, 0, 0)$, $(0, 0, 1, 0)$, $(0, 0, 0, 1)$ by their stages. What needs to be emphasized is that while training RBF neural networks, the k -means clustering algorithm is performed to determine the centers of hidden layers. c_i in Figure 4 is a hyperparameter that needs to be tuned. Back propagation and gradient descent are performed to obtain good classification ability. Finally, the output of SVM groups and neural networks are reshaped into one vector, which is used as the input of the exponential linear regression model. Artificial labels y are added to train the ELR model. The loss function of ELR is selected as mean square loss, namely,

$$L(w, b) = \frac{1}{m} \sum_{i=1}^m (EM_i - EM'_i)^2, \quad (5)$$

where EM_i is the evaluation of the i th patient's tumor malignance and EM'_i is the manually set supervising value.

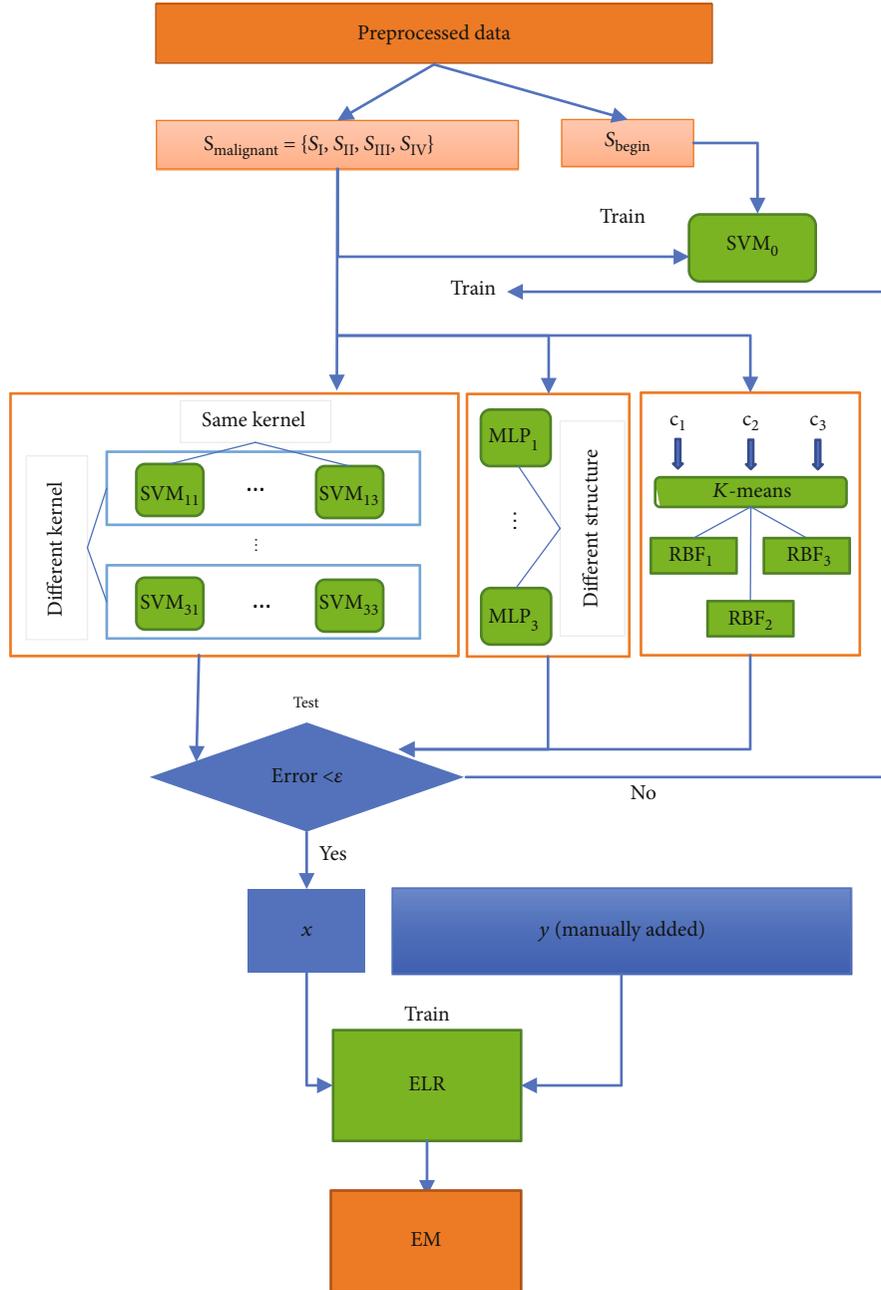


FIGURE 4: Training process of the proposed system.

TABLE 4: EM value of each stage of PCa.

Clinical stage of PCa	Range of ln EM
Stage I	2.7-3.6
Stage II	3.6-4.5
Stage III	4.5-5.3
Stage IV	>5.3

4.2. Analysis of the Results of Experiments. After the model was trained, all the malignant examples in different stages were input into the model and the range of their EM values was calculated, which are listed in Table 4. From Table 4, it

can be known that the EM values of all malignant examples have a rough 0.5 deviation around the supervising value set in advance. The model has good fitting ability on malignant samples of different stages, which indirectly proves our hypothesis that the tumor marker level increases exponentially with the development of tumor is credible.

To verify the effectiveness of our medical decision system, we compared the accuracy of the model on different scale datasets with the accuracy of doctors. As shown in Figure 5, when the amount of data is small, the accuracy of the auxiliary medical decision system is very low, close to 50%. In this circumstance, the accuracy of doctors is really high, almost 100%. However, as the amount of data increases,

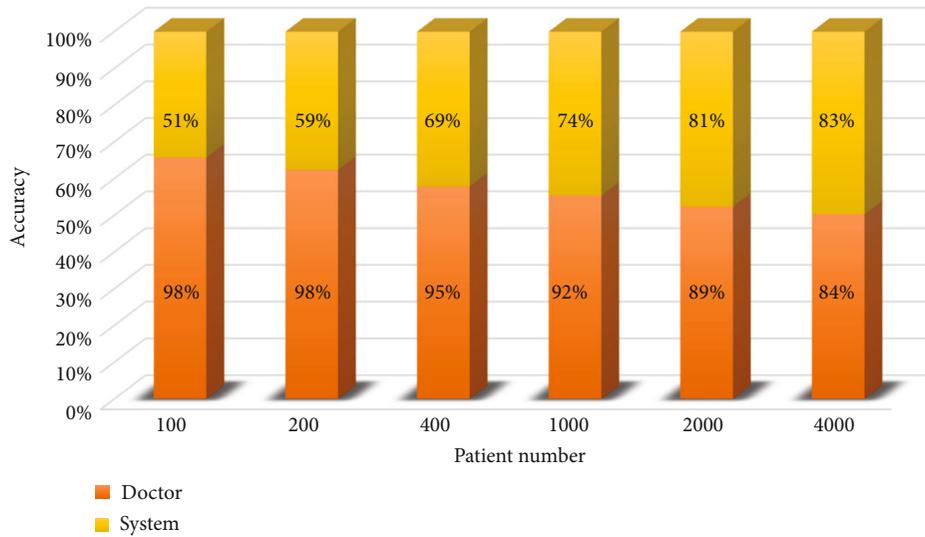


FIGURE 5: Comparison of the doctor and the system.

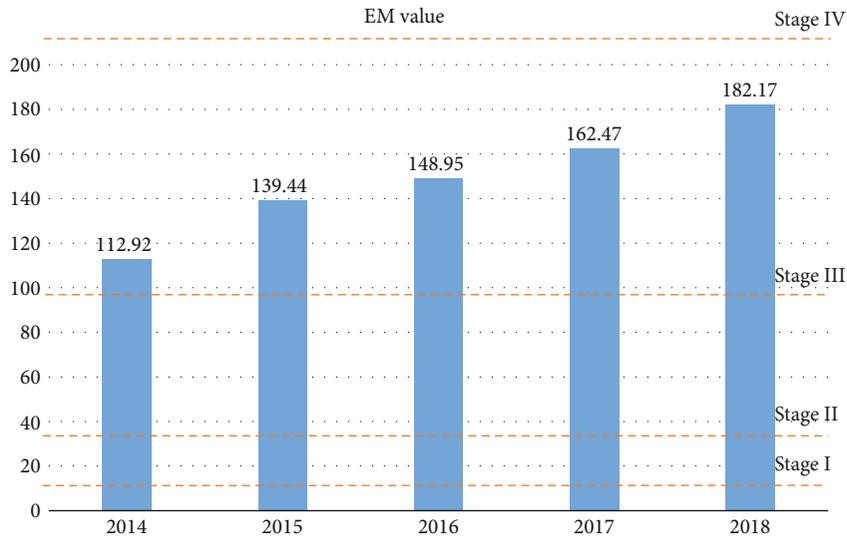


FIGURE 6: Average EM value in the past five years.

the accuracy of the medical decision system increases as well. Simultaneously, doctors' accuracy starts to decline because of the burden and cumulative errors. When the amount of data reaches 4000, the accuracy of the system is roughly the same with that of doctors. This indicates that our auxiliary diagnostic system can make use of the increasing amount of data to improve generalization performance.

We also calculated the average EM value of different years to explore the development trend of PCa in recent years. As shown in Figure 6, the mean EM value of patients from three hospitals has been gradually increasing since 2014. This implies an increase in the number or proportion of patients with malignant prostate cancer which will make medical resources scarcer, so it is necessary and urgent to establish an auxiliary medical decision system based on big data.

Because our medical decision system can quantitatively evaluate the malignancy of prostate cancer, it can easily judge

the efficacy of the treatment plan by its EM value change and recommend treatment methods to improve the condition of PCa patients according to their EM levels. Figure 7 shows the recommended treatment methods and changes of the EM value of a patient whose EM value is very high at first. In the end of the diagnosis interval, the patient's EM value is relatively low, which proves the tumor has been controlled by the recommended treatment plan. It can be concluded that the treatment methods recommended by the system can effectively improve the condition of cancer patients and prolong the survival time for patients in stage III or IV.

4.3. Relevant Analysis Based on the System. Since our medical decision system can evaluate the malignancy of tumors, by controlling different input variables, the influence of a certain factor on prostate cancer can be effectively evaluated. Here, relevant information of some patients was collated. Then,

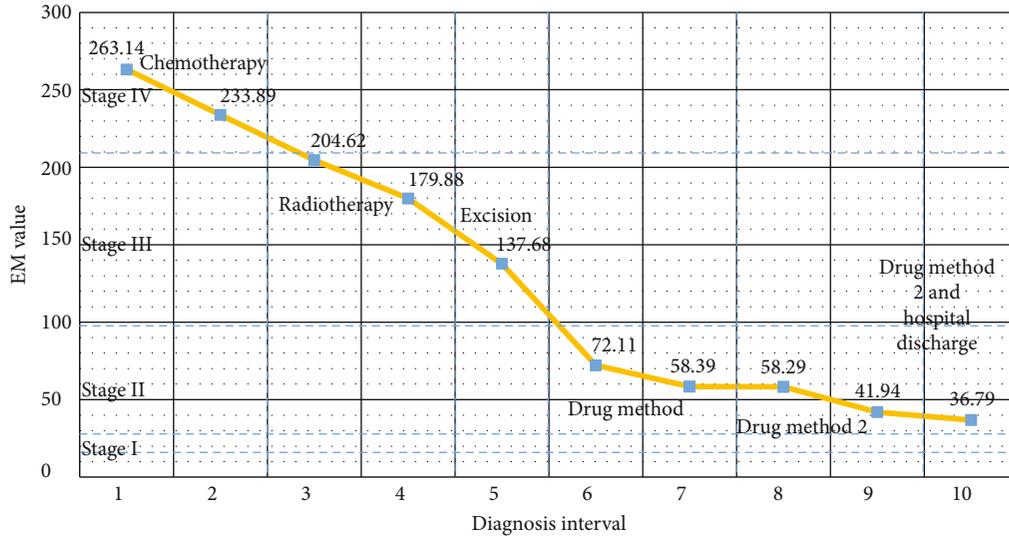


FIGURE 7: A typical treatment process of a PCa patient.

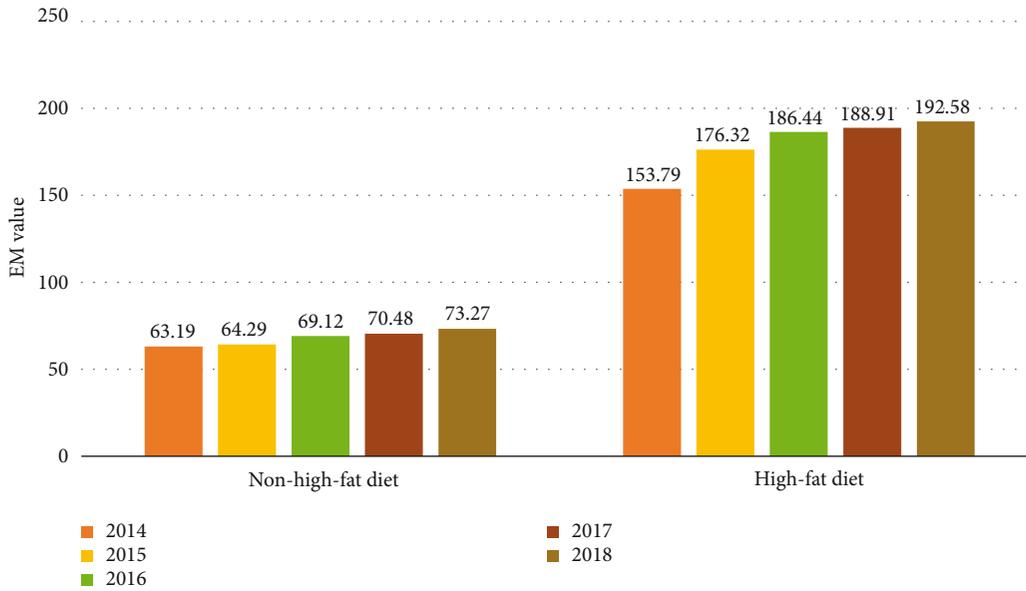


FIGURE 8: Contrast of people with different diet habits.

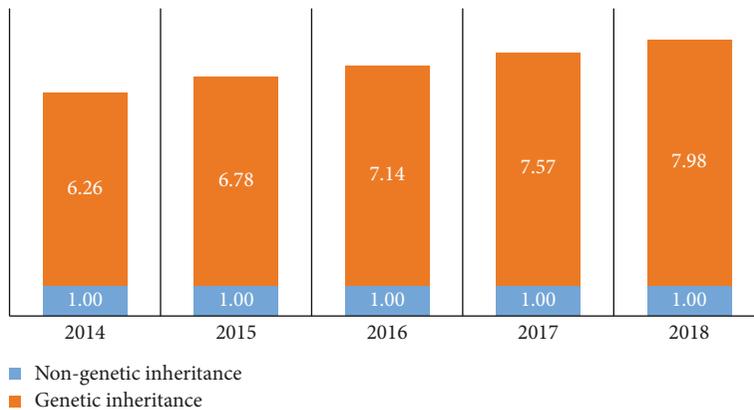


FIGURE 9: Contrast of people with or without genetic inheritance.

influence of patients' diet habits and genetic inheritance on prostate cancer was evaluated. Diet habits are mainly divided into high-fat diets and non-high-fat diets according to the description in the patient's medical history. From the data of 2014-2018, it can be seen that the condition of patients with high-fat diet tends to be more severe. The EM value for patients with high-fat diet is in the range of 150-190, while for those with non-high-fat diet, EM value is only 60-70, as shown in Figure 8.

Genetic inheritance is defined by a cancer case in the patient's family members. The results show that patients with genetic inheritance have a cancer malignancy that is 6 to 7 times that of patients without genetic inheritance, which can be seen in Figure 9.

5. Conclusion

This paper mainly builds an auxiliary medical decision system of PCa for developing countries that lack medical resources. The system is able to provide doctors with advice on the diagnosis, staging, and treatment method of prostate cancer. After training the system in big data environment, although its accuracy continues to rise, it still cannot replace professional doctors and can only be used as an auxiliary diagnostic system to relieve the burden of doctors. Based on this system, we have researched the development of prostate cancer in the past five years and found that the prevalence of prostate cancer is increasing. In addition, high-fat diet and genetic inheritance increase the severity of the disease. The next stage of this research will consider introducing other medical detection information, such as CT, MRI, and PET-CT, to further improve the accuracy and credibility of the system.

Data Availability

All medical data management and system come from Central South University. If readers are interested in those data, you can visit <http://www.xiangya.com.cn/english/>. All data analyzed during the current study are included in the submission.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

- [1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 394–424, 2018.
- [2] M. C. S. Wong, W. B. Goggins, H. H. X. Wang et al., "Global incidence and mortality for prostate cancer: analysis of temporal patterns and trends in 36 countries," *European Urology*, vol. 70, no. 5, pp. 862–874, 2016.
- [3] J. Ferlay, I. Soerjomataram, R. Dikshit et al., "Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012," *International Journal of Cancer*, vol. 136, no. 5, pp. E359–E386, 2015.
- [4] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, "Global cancer statistics," *CA: a Cancer Journal for Clinicians*, vol. 61, no. 2, pp. 69–90, 2011.
- [5] J. Wu and Z. Chen, "Data decision and transmission based on mobile data health records on sensor devices in wireless networks," *Wireless Personal Communications*, vol. 90, no. 4, pp. 2073–2087, 2016.
- [6] J. Wu, X. Tian, and Y. Tan, "Hospital evaluation mechanism based on mobile health for IoT system in social networks," *Computers in Biology and Medicine*, vol. 109, pp. 138–147, 2019.
- [7] J. Wu, Y. Tan, Z. Chen, and M. Zhao, "Decision based on big data research for non-small cell lung cancer in medical artificial system in developing country," *Computer Methods and Programs in Biomedicine*, vol. 159, pp. 87–101, 2018.
- [8] J. Wu, P. Guan, and Y. Tan, "Diagnosis and data probability decision based on non-small cell lung cancer in medical system," *IEEE Access*, vol. 7, pp. 44851–44861, 2019.
- [9] J. Wu, Y. Tan, Z. Chen, and M. Zhao, "Data decision and drug therapy based on non-small cell lung cancer in a big data medical system in developing countries," *Symmetry*, vol. 10, no. 5, p. 152, 2018.
- [10] B. Malmir, M. Amini, and S. I. Chang, "A medical decision support system for disease diagnosis under uncertainty," *Expert Systems with Applications*, vol. 88, pp. 95–108, 2017.
- [11] J. Wu, Z. Chen, and M. Zhao, "Community recombination and duplication node traverse algorithm in opportunistic social networks," *Peer-to-Peer Networking and Applications*, pp. 1–8, 2020.
- [12] J. Wu, Z. Chen, and M. Zhao, "An efficient data packet iteration and transmission algorithm in opportunistic social networks," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–17, 2019.
- [13] A. Tashkandi, I. Wiese, and L. Wiese, "Efficient in-database patient similarity analysis for personalized medical decision support systems," *Big Data Research*, vol. 13, pp. 52–64, 2018.
- [14] T. Kobayashi, "A blood tumor marker combination assay produces high sensitivity and specificity for cancer according to the natural history," *Cancer Medicine*, vol. 7, no. 3, pp. 549–556, 2018.
- [15] S. Saito, K. Taguchi, N. Nishimura et al., "Clinical usefulness of computer-assisted diagnosis using combination assay of tumor markers for pancreatic carcinoma," *Cancer*, vol. 72, no. 2, pp. 381–388, 1993.
- [16] B. K. Singh, "Determining relevant biomarkers for prediction of breast cancer using anthropometric and clinical features: a comparative investigation in machine learning paradigm," *Biocybernetics and Biomedical Engineering*, vol. 39, no. 2, pp. 393–409, 2019.
- [17] M. López-Riera, I. Conde, G. Quintas et al., "Non-invasive prediction of NAFLD severity: a comprehensive, independent validation of previously postulated serum microRNA biomarkers," *Scientific Reports*, vol. 8, no. 1, pp. 10606–10615, 2018.
- [18] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *COLT '92: Proceedings of the fifth annual workshop on Computational learning theory*, vol. 5, pp. 144–152, ACM Press, 1992.
- [19] R. Vijayarajeswari, P. Parthasarathy, S. Vivekanandan, and A. A. Basha, "Classification of mammogram for early detection

- of breast cancer using SVM classifier and Hough transform,” *Measurement*, vol. 146, pp. 800–805, 2019.
- [20] H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, “Using machine learning algorithms for breast cancer risk prediction and diagnosis,” *Procedia Computer Science*, vol. 83, pp. 1064–1069, 2016.
- [21] F. A. Badria, N. Shoaip, M. Elmogy, A. M. Riad, and H. Zaghloul, “A framework for ovarian cancer diagnosis based on amino acids using fuzzy-rough sets with SVM,” in *Advanced Machine Learning Technologies and Applications. AMLTA 2014. Communications in Computer and Information Science*, vol. 488, A. E. Hassanien, M. F. Tolba, and A. Taher Azar, Eds., pp. 389–400, Springer, Cham, 2014.
- [22] 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.
- [23] M. Abdar and V. Makarenkov, “CWV-BANN-SVM ensemble learning classifier for an accurate diagnosis of breast cancer,” *Measurement*, vol. 146, pp. 557–570, 2019.
- [24] A. Awad, M. Bader-El-Den, J. McNicholas, and J. Briggs, “Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach,” *International Journal of Medical Informatics*, vol. 108, pp. 185–195, 2017.
- [25] R. Khasha, M. M. Sepehri, and S. A. Mahdavian, “An ensemble learning method for asthma control level detection with leveraging medical knowledge-based classifier and supervised learning,” *Journal of Medical Systems*, vol. 43, no. 6, 2019.
- [26] R. Punmiya and S. Choe, “Energy theft detection using gradient boosting theft detector with feature engineering-based pre-processing,” *IEEE Transactions on Smart Grid*, vol. 10, no. 2, pp. 2326–2329, 2019.
- [27] G. Li, G. Zhao, C. Zhou, and M. Ren, “Stochastic Elastic Properties of Composite Matrix Material with Random Voids Based on Radial Basis Function Network,” *International Journal of Computational Methods*, vol. 15, no. 1, article 1750082, 2017.