A Medical Support System for Prostate Cancer Based on Ensemble Method in Developing Countries

QingHe Zhuang[1,2], Jia Wu[1,2][0000−0001−9013−0818], and GengHua Yu[1,2]

[1] School of Computer Science, Central South Universtiy, Changsha, China
[2] "Mobile Health" Ministry of Education-China Mobile Joint Laboratory, Chang sha, China

Abstract. As a cancer with high incidence rate, Prostate cancer (PCa) endangers men's health worldwide. In developing countries, medical staff are overloaded because of the lack of medical resources. Medical support system is a good technique to ease contradiction between the large number of patients and small number of doctors. In this paper, we have collected 1,933,535 patient information items from three hospitals, constructed a medical support system for PCa. It uses six relevant tumor markers as the input features and output a quantitative indicator EM value for staging and recommending treatment method. Classical machine learning techniques, data fusion and ensemble method are employed in the system to make the results more correct. In terms of staging PCa, it reaches an accuracy of 83%. Further research based on the system and collected data are carried out. It is found that the incidence of prostate cancer has been rising in the past five years and diet habit and genetic inheritance have a great impact on it.

Keywords: prostate cancer· tumor marker· medical support system· machine learning.

1 Introduction

In 2018, PCa's morbidity and mortality are 13.5% and 6.7% respectively in male patients. In 185 countries, it 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.

Even though PCa is not high-fatal, in developing countries that lack medical resources, many patients can't receive timely and effective diagnosis and therapy, which will worsen the condition of patients. Scarce medical resource, specially the lack of high-quality medical resources may lead to patients' distrust to doctors and aggregate the conflict between them[2]. Sometimes doctors even get physically injured by family members of patients because of their distrust to doctors. Take China for example, there are only 2.59 practitioners for every 1000 people[3]. Patients need to wait hours for diagnosis or examination[4]. Things get worse in some top-class hospitals. In Beijing, a small number of medical staff

in 3A grade hospitals need to serve not only over 20 million local people but also people seeking treatment from other regions. Overloaded burden increases the chance of mistakes and causes severe consequences. At the same time, inspection charge of PCa accounts for a large portion. Many high-end inspection methods including MRI and PET-CT are too expensive for poor patients to afford.

Other developing countries may face similar dilemmas:

• Due to the scarce medical resources, it is difficult for patients to get timely diagnosis and treatment.

• The long-term workload of doctors increases the chance of mistakes and aggregates the conflicts between doctors and patients.

• Many hospitals in developing countries have poor medical equipment and many patients in developing countries cannot afford expensive checking fare.

Scarce medical resources, long-term overloaded medical staff and difficult access to medical care have severely restricted the life expectancy of patients in developing countries. Fortunately, these problems may be eased by building medical support system which aims to offer help for medical staff[4]. By analyzing large amounts of data, the medical support system can extract a diagnostic model. It will serve doctors with suggestions relevant to diagnosis or treatment based on the learned model[6][7]. Combining suggestions form the system and their own knowledge, doctors will give the final decision. Contradiction between doctors and patients may be eased if the medical support system work well[8][9]. In this work, we constructed a medical support system which can diagnose if a patient has PCa, determine the pathological stage, recommend treatment method, and evaluate the effectiveness of treatment method. Given the low income in developing countries, the system is featured as inputting six tumor markers with relatively low testing price and high relevance to PCa. Classical machine learning techniques and ensemble method are adopted to extract the knowledge inside data and improve system's performance.

Compared with other works, the main innovations and contributions of our system include:

• Unlike other system based on CT or MRI, the constructed system only uses features that are cheap for people in developing countries and also reached good performance, which makes it possible to deploy the system in other developing countries.

• In addition to the diagnosis, the system can give treatment plan and evaluate its effectiveness quantificationally at the same time.

• The system is trained based on a large amount of patient information from three high-level hospitals in China and some factors affecting PCa are analyzed via the constructed system.

2 Design of Medical Support System

2.1 Overall Requirements and Framework of the System

Medical support system aims to provide some advice for doctors. Its functions contain determining PCa's benignancy or malignancy and pathological stage,

recommending suitable treatment plan and judging effectiveness of treatment plan. Determining PCa's benignancy or malignancy and pathological stage can be regard as a classification task. But in order to give a cancer treatment plan and estimate its efficacy at the same time, the whole problem is considered as a regression problem. Our system will transform the discrete input feature into continuous variable which cane evaluating the malignancy of PCa, abbreviated for EM. The larger EM stands for the higher malignancy. If the value does not decrease after a certain treatment plan is executed, it means that the treatment plan has little efficacy and another a new one needs to be selected. At the same time, the medical support system needs to have good parallelism and be able to process multiple patients' simultaneous diagnosis requests. After the medical system is invested, the amount of data obtained will increase over time. The decision model will be retrained to further improve the generalization performance.

2.2 Detailed Description of the Medical Support System

In this medical support system, six tumor markers with high relevancy to PCa is used as input feature. They include 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). Machine learning techniques including Support Vector Machines (SVM) and Neural Networks (NN), mainly Multilayer Perceptron (MLP) and Radical Basis Function Neural Network (RBF) are ensembled to acquire good performance in classification.

Fig. 1 depicts the main flow of the medical support system. First, in preprocessing module, relevant data from different hospital systems are collected. Then six important tumor markers' level are extracted from thousands of information items. After data cleaning and normalizing, input vector $x = (x_{PSA}, x_{PSMA}, x_{tPSA}, x_{RBC}, x_{HB}, x_{PAP})$ is formed. In decision module, it will firstly use a binary SVM to determine tumor's malignancy or benignancy. In clinical experience, increase of tumor marker doesn't mean a malignant tumor for sure. They don't have high specificity and there may be other reasons that lead to the increase such as lesions or inflammations. So the system cannot determine it simply by critical threshold. Sometimes one tumor marker is abnormal while others are normal. In this circumstance, doctors will find it hard to give correct results. Data mining or machine learning models are able to extract features with high specificity and make use of these features for decision. That's the motivation they are deployed in the system. If diagnosed as benign tumor, relevant therapy will be recommended according to the previously recorded similar samples in the database. If judged as malignant tumor, then an ensemble model is executed to complete pathological stage division. The pathological stage for malignant tumor includes four stages: I, II, III, and IV. That is to say, the system must complete a four-classification task. Since SVM is mainly used in binary classification, multi-classification SVM (M-SVM) which combines binary SVM and DS (Dempster/Shafer) evidence theory is constructed according to [10]. Unlike binary SVM, M-SVM can output probabilistic result and reduce the number of

used binary SVMs compared with other extending method. The output result from a M-SVM is a four-dimensional vector, whose value in each dimension represents the possibility of corresponding class. In order to reduce the risk brought by choosing specific kernel function, different M-SVMs with different kernels are trained as is shown in Fig. 1. Here three commonly used kernel functions: linear kernel, polynomial kernel and Gaussian kernel are chosen simultaneously. For binary SVMs in each M-SVM, the same kernel functions are used. While in different M-SVM, different kernel functions are used.

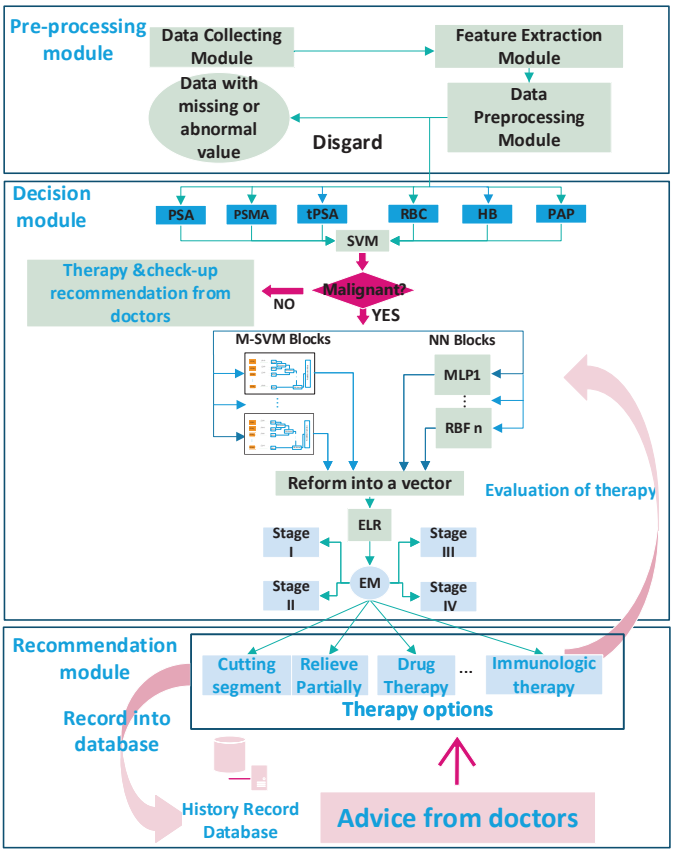

Fig. 1. A figure caption is always placed below the illustration. Please note that short captions are centered, while long ones are justified by the macro package automatically.

While training, hyperparameters in kernel functions and binary SVMs are tuned to reduce the generalization error below threshold ϵ. In order to reduce risk further, widely used MLP neural network and RBF neural network are added into the system. Because 6 input features are chosen and the 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. ReLU function is used as the activation function in MLP neural networks. Similarly, three RBF networks with different structure are used. The hidden unit number in three networks are set as 10, 14, 16, respectively. Use k-means clustering algorithm to determine the center c_i of each hidden unit. In RBF neural network, radical basis function is used as the activation function. As in SVMs, the hyperparameters are adjusted to reduce the generalization error below the threshold ϵ.

Ensemble algorithm

Input:
Training set: $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}, y_i \in \{I, II, III, IV\}$.
Primary classifier: $S = \{SVM_1, \ldots, RBF_3\}$
Output: Second learning algorithm $H(x) : ln(y_{EM}) = w^T x + b$
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
  $y'_i = map(y_i)$; /*map function convert the class label into a numerical value.*/
  $D' = D' \cup ((z_{i1}, z_{i2} \ldots z_{it}), y'_i)$
end for
use D' to train $H(x)$;
output $H(x)$;
End

Finally, outputs of each M-SVM and all MLP and RBF networks are reshaped into one vector as the input of the secondary learner. Instead of averaging the results of different classifiers, a second learner is introduced to learn the weight of each base learner. The selection of the second learner is based on the priori knowledge. By observing the tumor marker level of all samples, it is found that for benign tumors and patients in stage I, the tumor marker levels are usually close to normal range. But for patients in stage III and stage IV, the level of tumor markers deviates greatly. Therefore, we made an priori assumption that the increase of tumor markers conforms to the exponential law as PCa worsens. This hypothesis is basically true in medicine. In the early stages like stage I and stage II, symptoms are slight or not obvious. Tumors are often latent and grow slowly. However, in stage III or IV, they develop savagely and spread throughout the body, causing the tumor marker levels really high. Hence, Exponential Linear Regression (ELR) is selected as a secondary learner to ensemble the results of M-SVM, MLP and RBF models. We add supervising label 3,4,5,6 manually for the input patient samples in stage I II III IV, respectively. The output value of

exponential linear regression model is not set to start from 1 in order to improve the model's robustness to normal people and benign tumor cases. Finally, the evaluation of PCa's malignancy (EM) is output. Ensemble algorithm shows the procedure integrating the results of base learners by exponential linear regression. Pathological stage of malignant PCa is determined by EM value.

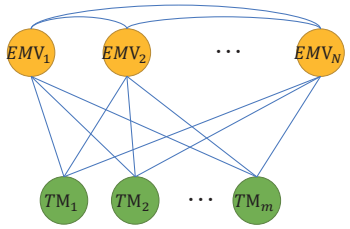

Fig. 2. Relationship between EMV and TM.

In recommendation module, a probabilistic graphical model on the basis of that treatment selection mainly depends on EM value is used. The relationship between EM value (EMV) and treatment method (TM) is shown in Fig. 2. Selection of TM is related to EMV. Meanwhile, selection of TM will also have an impact on EMV. They depend on each other. Fig. 3 depicts the building and training process of the model. First, numerical EM value is divided by interval parameter e to form discrete set of EMV ($EMVS = emv_1, emv_2, \ldots, emv_N$) and treatment method set (TMS) containing commonly used tumor treatment methods such as chemotherapy, radiotherapy, excision, drug method, hospital charge is formed simultaneously. Treatment process of patient i can be characterized as a sequential data: $\left(E_i^1,\ T_i^1\right),\ \ldots, (E_i^t,\ T_i^t),\ \ldots, \left(E_i^{|t_i|},\ T_i^{|t_i|}\right),\ i = 1, 2, \ldots Tot$ where E_i^t and T_i^t represents EM value and treatment method of patient i in treatment interval t and belong to EMVS and TMS respectively. $|t_i|$ is the total treatment interval of patient i, Tot is total number of data.

Use these data, two kinds of conditional probability distribution, $P(EMV|TM, EMV)$ and $P(TM|EMV)$ can be learned. $P(TM|EMV)$ means the possibility of selecting TM in the state of EMV. Similarly, $P\left(EMV|TM, EMV\right)$ means the possibility of EM value changing into state EMV after TM is used in state EMV. Parameter can be estimated by (1) and (2).

$$P\left(TM = tm_k|EMV = emv_j\right) = \frac{1 + \sum_{i=1}^{Tot}\sum_{t=1}^{|t_i|} sgn\left(E_i^t = emv_j, T_i^t = tm_k\right)}{N + \sum_{i=1}^{Tot}\sum_{t=1}^{|t_i|} sgn\left(E_i^t = emv_j\right)} \tag{1}$$

$$P\left(EMV = emv_m|TM = tm_k, EMV = emv_j\right) = \frac{1 + \sum_{i=1}^{Tot}\sum_{t=1}^{|t_i|-1} sgn\left(E_i^t = emv_j, T_i^t = tm_k, E_i^{t+1} = emv_m\right)}{NM + \sum_{i=1}^{Tot}\sum_{t=1}^{|t_i|} sgn\left(E_i^t = emv_j, T_i^t = tm_k\right)} \tag{2}$$

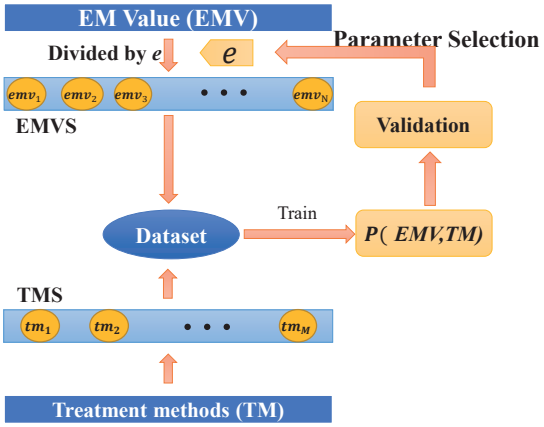

Fig. 3. Building and training process of the recommendation model.

Selection of interval e follows (3):

$$e = argmax_e \prod_{i=1}^{Tot} P\left(E_i^1\right) P\left(T_i^{|t_i|} | E_i^{|t_i|}\right) \prod_{t=1}^{|t_i|-1} P(T_i^t \left| E_i^t\right) P(E_i^{t+1} \left| E_i^t, T_i^t\right) \quad (3)$$

When new patient with malignant PCa comes, the system will recommend treatment method to doctors. The recommendation procedure is rough and primary so final decision must be made by doctors. But it can relieve burden of medical staff to some extent.

3 Experiment

3.1 Dataset and Models' Training

From three top-class hospitals in China: First Xiangya Hospital, Second Xiangya Hospital and Third Xiangya Hospital, we obtained a large amount of data. Records of the tumor markers (PSA, PSMA, tPSA, RBC, HB, and PAP), diagnostic results (benign, stage I, stage II, stage III, stage IV) and treatment process are screened and preprocessed. A total of 12186 patients' data is extracted and the distribution is shown in Table 1.

Table 1. Distribution of the collected data.

	Benign	Malignant			
		Stage I	Stage II	Stage III	Stage IV
Number	3628	2864	1523	1795	2376

Table 2 shows the normal range of tumor markers relevant to PCa. Values of malignant patient's tumor marker are several times or even tens of times beyond the normal range.

Table 2. Normal range of different tumor marker.

Types of tumor marker	Normal range
Total Prostate-Specific Antigen	4-20 $\mu g/L$
Hemoglobin	120-165 $\mu g/L$
Red Blood Cell	12-15 $g/100ml$
Prostate Acid phosphatase	0-9 U/L
Prostate-specific membrane antigen	0-4 ng/mL

The data set is divided into three parts: training set, validation set and test set, accounting for 70%, 20% and 10% respectively. First of all, we choose the appropriate kernel function and penalty parameter to train SVM_0. Secondly, we extract all malignant samples for the training in the next step. Because SVM and neural networks are not sensitive to data, and arbitrary division of data is likely to lead to the problem of imbalanced data which means two datasets don't have same distribution. Hence we choose the same training set to train all base learners. For M-SVM blocks' training, malignant samples are firstly divided in two ways which have been described elaborately in [10], forming two datasets $S_A = \{S_{I,II}, S_{III,IV}\}$ and $S_B = S_{I,III}, S_{II,IV}$, where $S_{i,j}$ represents the union of $c_i,\ c_j$. S_A and S_B are then used to train two binary SVMs. The output of M-SVM is the fusion result of these two binary SVMs. For neural networks, the malignant samples are directly marked as $(1,0,0,0)^T, (0,1,0,0)^T, (0,0,1,0)^T, (0,0,0,1)^T$ by their stages. Back propagation and gradient descent are performed to obtain good classification ability. Finally, the output of M-SVM blocks and neural networks are reshaped into one vector, which is used as the input of exponential linear regression model. The loss function of exponential linear regression uses mean square loss function.

3.2 Analysis of the results of experiments

After training, all malignant samples are input into the model, and calculate the range of their EM values, which are listed in Table 3. As Table 3 shows, EM values of all the malignant examples have a rough 0.5 deviation around the supervising value set in advance. The model has a good fitting ability on the malignant samples of different stages, which indirectly proves the correctness of our hypothesis that the tumor marker level increases exponentially as PCa worsens.

In order to verify the effectiveness of our medical support system, accuracy of the system on different scale data sets with that of doctors and other method are compared. As shown in Fig. 4, when the dataset is small, accuracy of medical

Table 3. EM value of each stage of PCa.

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

support system is very low, no better than random guess (50%). While accuracy of doctors is high, almost 100%. However, when the size of dataset increases, accuracy of the medical support system increases and doctors' accuracy starts to decrease because of overloaded burden and cumulative errors. As the amount of data reaches 4000, accuracy of the system is roughly the same with that of doctors. This implies our medical support system can make full use of the increasing amount of data to improve generalization performance.The final result is based on 12186 patients' data. For SVM_0, it reaches 89.9%, 86.6%, 98.9% in terms of accuracy, recall and precision (take malignant samples as positive). In Fig. 4, it can also be seen that accuracy of proposed model is always higher than others' instead of being an average of other models. This demonstrates the correctness of strategy of using secondary learners to ensemble them. It may be explained as that base classifiers play a role of feature extractor and this makes it easier for secondary learner to determine the clinical stage. The confusion table of the proposed ensemble model is shown in Table 4. The final accuracy is roughly 87.4%.

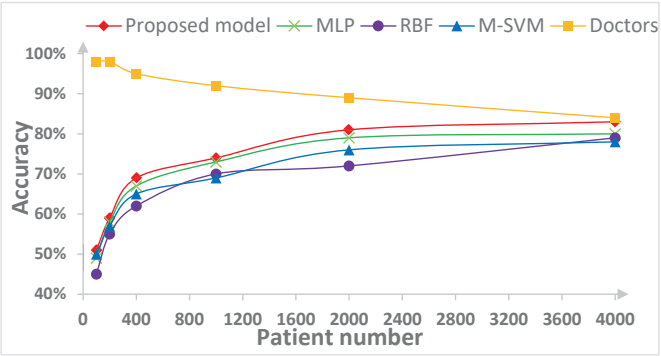

Fig. 4. Comparison of doctors and the system.

Average EM value of different years are calculated to explore the development trend of PCa in recent years. As shown in Fig. 5, the mean EM value of patients from three hospitals has been increasing from 2014 to 2018. This implies an increase in proportion of patients with PCa which will make medical resources

Table 4. Confusion Table of the ensemble model.

Prediction	Real label			
	Stage I	Stage II	Stage III	Stage IV
Stage I	2621	101	24	9
Stage II	204	1312	67	25
Stage III	25	66	1476	271
Stage IV	14	44	228	2071

more precious, so it is necessary and urgent to establish a medical support system.

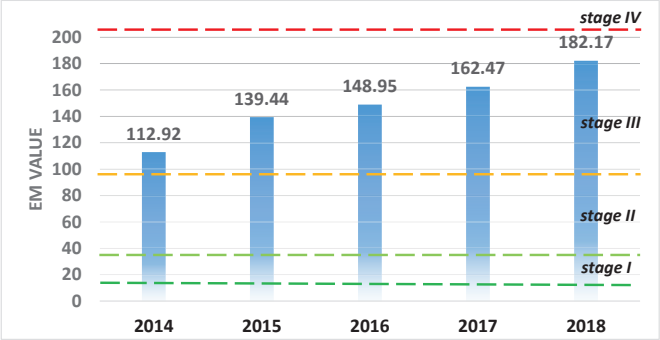

Fig. 5. Average EM value in the past five year.

By calculating the quantitative indicator of PCa's malignancy (EM value), the system can easily judge therapy's efficacy by the change of EM and recommend treatment methods to improve the condition of PCa patients. Fig. 6 shows the recommended treatment methods and changes of EM value of one patient diagnosed with stage IV PCa. At first, EM value is very high. In the end of several diagnosis intervals, the patient's EM value comes to a relatively low level, which proves the tumor has been controlled by the recommended treatment plan. It is convincing that the system can make some decisions and relieve doctor's burden indeed.

3.3 Relevant analysis based on the system

By controlling different input variables, influence of a certain factor on prostate cancer are evaluated. Here, relevant information of some patients is collated to evaluate impact of patients' diet habits and genetic inheritance. Diet habits are divided into high-fat diets and none high-fat diets by description in medical record. Genetic inheritance is defined by malignant tumor incidence in the patient's family members (lineal relatives). From the data of 2014-2018, it can be

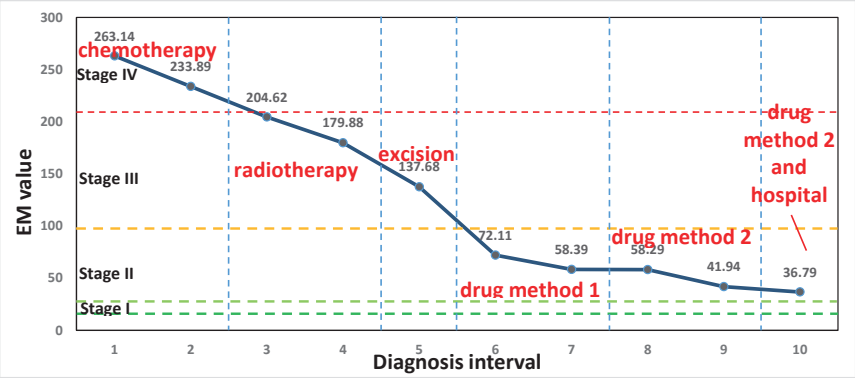

Fig. 6. A treatment process of a PCa patient.

inferred that high-fat diet with tends to worsen the condition of patients. EM value for patients with high-fat diet is 2.43-2.63 times of those without high-fat diet, and in terms of genetic inheritance, it's 6.26-7.98 times, as shown in Fig. 7.

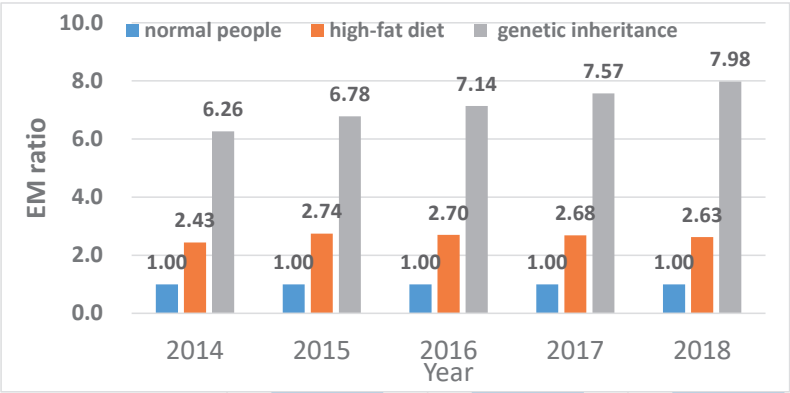

Fig. 7. Influence of diet habit and genetic inheritance.

4 Conclusion

This paper mainly builds medical support system of PCa for countries that lack medical resources. The selected features are cheap for underdeveloped countries, which ensures that even poor people have access to cancer health care. The system is able to provide doctors with advice on the diagnosis, staging and therapy recommendation of PCa. After training the system in the big data environment,

it gets relatively good results. It can relieve the burden of doctors to some degree but can't replace the doctor completely. In many cases, it needs doctor's correction. Using the system, development of prostate cancer in the past five years is researched, and found the increasing prevalence of PCa, which proves the significance of establishing the medical support system. In addition, high-fat diet and genetic inheritance increase the severity of the disease.

References

1. B. Freddie, F. Jacques, and S. Isabelle, "Global cancer statistics 2018:GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA. Cancer J. Clin., vol. 68, no. 6, pp. 394–424, 2018.
2. 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, no. November, pp. 44851–44861, 2019.
3. Department of Planning Development and Information Technology, "Statistical Communiqué on China's Health Care Development in 2018," 2019.
4. J. Wu, X. Tian, and Y. Tan, "Hospital evaluation mechanism based on mobile health for IoT system in social networks," Comput. Biol. Med., vol. 109, no. April, pp. 138–147, 2019.
5. 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 (Basel)., vol. 10, no. 5, pp. 1–16, 2018.
6. B. Malmir, M. Amini, and S. I. Chang, "A medical decision support system for disease diagnosis under uncertainty," Expert Syst. Appl., vol. 88, pp. 95–108, 2017.
7. C. D. Stylios, V. C. Georgopoulos, G. A. Malandraki, and S. Chouliara, "Fuzzy cognitive map architectures for medical decision support systems," Appl. Soft Comput. J., vol. 8, no. 3, pp. 1243–1251, 2008.
8. P. Wang, P. Zhang, and Z. Li, "A three-way decision method based on Gaussian kernel in a hybrid information system with images: An application in medical diagnosis," Appl. Soft Comput. J., vol. 77, pp. 734–749, 2019.
9. A. Tashkandi, I. Wiese, and L. Wiese, "Efficient In-Database Patient Similarity Analysis for Personalized Medical Decision Support Systems," Big Data Res., vol. 13, pp. 52–64, 2018.
10. Z. Hao, L. Shaohong, and S. Jinping, "Unit model of binary SVM with DS output and its application in multi-class SVM," Proc. - 2011 4th Int. Symp. Comput. Intell. Des. Isc. 2011, vol. 1, no. 2, pp. 101–104, 2011.