IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 20, 2020, accepted September 15, 2020, date of publication September 18, 2020,
date of current version September 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3024840

Decision-Making Model Based on Ensemble Method in Auxiliary Medical System for Non-Small Cell Lung Cancer

HUANZE CHEN[1,2], WANGPING XIONG[3], JIA WU[1,2], (Member, IEEE), QINGHE ZHUANG[1,2], AND GENGHUA YU[1,2]

[1] School of Computer Science and Engineering, Central South University, Changsha 410083, China
[2] "Mobile Health" Ministry of Education—China Mobile Joint Laboratory, Changsha 410083, China
[3] School of Computer Science, Jiangxi University of Traditional Chinese Medicine, Nanchang 330006, China

Corresponding authors: Jia Wu (jiawu5110@163.com) and Wangping Xiong (20030730@jxutcm.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61672540 and Grant 61762051, in part by the Hunan Provincial Natural Science Foundation of China under Grant 2018JJ3299 and Grant 2018JJ3682, and in part by the Fundamental Research Funds for the Central Universities of Central South University under Grant 2019zzts152.

ABSTRACT In many countries, lung cancer is the chief cancer type, and the overall survival rate in 5 years remains at a low rate of 16.8%. Among lung cancer patients, the proportion of patients with non-small cell lung cancer (NSCLC) can reach 85%. Especially in developing countries, it has become a great challenge for doctors to diagnose efficiently because of the complex diagnostic process of NSCLC, a large number of patients, and limited medical resources. Based on the above situation, our main objective in this study is to construct an auxiliary Artificial Intelligence medical system for doctors to diagnose the NSCLC patients efficiently. In this research, we select the combination of Support Vector Machine (SVM) and Artificial Neural Network (ANN) to complete the classification and training tasks of the medical system. In this study, we trained and tested our decision model based on the data information of 12,186 patients in three hospitals in China. The results of this experiment show that the accuracy of our system has reached 88% when the amount of data reaches 4000, which is close to that of doctors.

INDEX TERMS Big data, diagnostic parameters, machine learning model, medical decision making, NSCLC.

I. INTRODUCTION

According to the cancer report of GLOBOCAN 2018, there were approximately 18 million new cancer cases and 9.6 million deaths all over the world in 2018. Lung cancer is the chief cancer type with the highest morbidity and mortality and non-small cell lung cancer accounts for approximately 85% of all lung cancer cases [1]–[3]. Additionally, 50% of NSCLC patients are diagnosed at advanced stages. For them, systemic chemotherapy is the recommended treatment, but the average survival time is only 7 to 11 months, and the 5-year survival rate is about 20%. However, if a NSCLC patient can be diagnosed and accept the appropriate treatment plan in the early stage, the 5-year survival rate will increase to 80%. Thus, an efficient method of determining the malignancy of the tumor and the clinical stage in the early stage is required urgently for the NSCLC patients.

The situation goes worse in most developing countries. Because of the huge population, a small number of medical staff, limited medical resources and equipment, most NSCLC patients could not timely receive the efficient diagnosis in their early stage and miss the best time for treatment [4]–[7]. When the tumor develops to an advanced stage, the patient has to receive treatments such as the radiotherapy and chemotherapy which will also bring a certain degree of pain and harm to the patient.

China, the largest developing country in the world, faces similar problems.

(1) There is an imbalance between the number of medical staff and patients. According to the statistics of the Chinese Ministry of Health in 2016, more than 5600 people share one doctor, and every doctor must take care of 72 patients a day.

The associate editor coordinating the review of this manuscript and approving it for publication was K. C. Santosh.

(2) Limited medical resources are unevenly distributed. Data from the Chinese Ministry of Health shows that 80% of the medical resources are distributed in large cities and developed areas but received by 7% of the population. The remaining 20% must be shared by 93% of the population.

(3) The diagnostic process of NSCLC is complicated. NSCLC patients usually require PET-CT tomography to detect the location and severity of the lesion. However, each time a PET-CT scan is performed, at least 640 images will be generated, requiring at least 2GB of storage memory space. As for doctors, it is inefficient and impractical to review all the images. Only part of the images can assist doctors in diagnosing, which is an excellent waste of medical resources.

(4) Patients have a tense relationship with doctors. Due to the short diagnosis time and patients' little medical knowledge, there is a crisis of confidence between doctors and patients. According to the report about the Doctor-patient relationship from the Chinese Medical Doctors Association (CMDA), 74.29% of doctors believe that their legitimate rights and interests cannot be protected, and 47.35% think that the current practice environment is poor.

These problems can be solved by establishing an Artificial Intelligence medical decision system. By analyzing data from NSCLC patients, we can build a diagnostic and recommended model [8], [9] using the tumor markers as the input features. For the patients with malignant tumor, the levels of tumor markers have a negative correlation with their survival time. And the main tumor markers which have a great correlation with the detection and diagnosis of NSCLC are soluble fragment of Cytokeratin (CYFRA21-1), Carcinoembryonic antigen (CEA) and Cancer antigen (CA)-125 [10]. However, it is actually a tedious work for doctors to tell the malignancy of NSCLC by the statistic report of tumor markers. Therefore, it is necessary to propose a auxiliary medical system for doctors. When a new patient arrives, the system will provide doctors with basic diagnosis results and treatment advice after the patient's physiological information is tested and analyzed by the system. Based on the above results and medical experience, the doctor will give the final diagnosis results and detailed treatment process, thereby improving the efficiency of diagnosis, reducing the rate of misdiagnosis and decreasing the workload of doctors. What need to be emphasized is that our medical system could not replace doctors to make the final decisions for patients although the accuracy of the system remains a high level.

This article constructs a decision model to diagnose the NSCLC patients and stage cancer based on SVM, ANN and the ensemble method. Specifically, we use a single SVM for the binary classification task of telling the tumor is benign or malignant. And a combination of SVM and ANN is adopted in our system to determine the clinical stage of the tumor, which is actually a four-classification problem. Moreover, a recommended system is designed to give the treatment plan and to evaluate the treatment plan. Considering the relative scarcity of medical resources in developing countries, we select three tumor markers which are cheap to test but have high relevance to NSCLC as the input features and use SVM and classic ANN as the basic algorithm. The contributions of the study include:

(1) With the analysis of the database of NSCLC patients, a medical auxiliary system for NSCLC is designed to give recommendations of treatment plan, which will reduce the workload of doctors and improve the diagnosis efficiency.

(2) Based on SVM, ANN and the ensemble learning, a decision model for NSCLC with high accuracy and stability is constructed to determine whether a patient suffers from NSCLC and the clinical stage of the patient's tumor.

(3) An evaluation standard for influential factors on NSCLC was proposed. In this study, two influential factors, diet habit and genetic inheritance, were tested. Similarly, we can use this standard to test other factors such as alcohol consumption.

(4) The training data is obtained from three high-level hospitals in China, and influential factors of NSCLC are analyzed through our medical system.

The remainder of this article is organized as follows: In Section II, we describes some excellent studies which is related to our work. In Section III, the design of the system model will be proposed. Moreover, all of the experimental results are shown in Section IV. In Section V, we describes the discussion of this article. And the last section is the conclusion of the paper.

II. RELATED WORK

In recent years, many Artificial Intelligence medical systems are applied in detecting, diagnosing, and treating cancer, which reflects that Artificial Intelligence medical systems have become hot research directions in the medical field.

Compared with the medical image method such as CT and PET-CT, using tumor markers to detect patients' tumor in a medical system is a relatively cheap method for developing countries with limited medical resources. Many studies [11]–[13] adopted different combinations of tumor markers to build the medical system. Specifically, Saito *et al.* [11] analyzed 17 tumor markers from 145 patients with pancreatic cancer and selected nine tumor markers for the medical system. The experimental results showed that this system could accurately distinguish the malignant pancreatic cancer from the benign one. Kobayashi [12] adopted the combination of growth-related tumor markers for primary cancer detection. The final results showed 80-90% sensitivity, 84-85% specificity, and 83-88% accuracy. Feng *et al.* [13] used the ANN model with six serum tumor markers and 19 parameters of lung cancer to distinguish lung cancer from benign lung disease. Based on the data of 117 lung cancer patients and 93 lung benign disease patients, the experimental results showed 98.3% sensitivity, 99.5% specificity, and 96.9% accuracy. However, most of studies which use the tumor markers as the input of the model focus on the functions of determining the tumor is benign or malignant and its clinical stage rather than the function of recommendations of treatment plan.

For the algorithm used in the medical system, both SVM and ANN are widely used [14]–[16] and show great performance in the classification problems. Specifically, Kureshi *et al.* [14] designed a predictive model for personalized therapeutic interventions in NSCLC using patient data. The frequent pattern mining was used to establish the relationship between patient characteristics and tumor in advanced NSCLC. This model used four classifiers including SVM, open source Java implementation of the C4.5 algorithm (J48), Random Forest and classification and regression tree (CART) to predict the treatment plan. Moreover, this model provided a framework to evaluate the new forms of phenotypic information and genomic data. The experimental results implied that the highest accuracy of the classifies was 76.56%. Pu *et al.* [15] demonstrated the performance of ANN was better than that of multiple linear regression model in overcoming the limitations of the international TNM staging standards for predicting the survival time of patients with NSCLC. The authors achieved better prognostication of survival by including additional prognostic factors. Prognostic reference factors in this study include FDG-PET measurements, treatment variables, and other clinical factors. Based on the data of 328 patients with NSCLC, the ANN reduced standard deviation from 17.4 months to 14 months, compared with the multiple linear regression model. Tirzīte *et al.* [16] applied SVM to distinguish patients with lung disease from healthy volunteers based on the analysis of exhaled breath with an artificial olfactory sensor. A Cyranose 320 sensor device was used for the collection, and the related breath data were analyzed by SVM. Finally, the SVM in this study could distinguish patients with lung cancer from healthy volunteers in 98.8% of cases.

For the ensemble learning, it is a widely used method which combines multiple models, such as classifiers or experts, to solve a computational intelligence problem. The ensemble learning can improve the performance of the total model, so it is also widely used in the area of medical auxiliary diagnosis. Award *et al.* [17] introduced a new machine learning based framework for Early Mortality Prediction for Intensive Care Unit patients (EMPICU). They employ the ensemble learning Random Forest (RF), the predictive Decision Trees (DT), the probabilistic Naive Bayes (NB) and the rule-based Projective Adaptive Resonance Theory (PART) models. And the experimental results showed the proposed EMPICU framework outperformed standard scoring systems in terms of AUROC and time. Abdar and Makarenkov [18] proposed a new ensemble-based classification model for breast cancer detection. The main objective of this study is to expand an automatic expert system to provide an accurate diagnosis of breast cancer. In this study, the data of breast cancer was analyzed with SVM and ANN and a boosting ensemble technique for the diagnosis of breast cancer was constructed. The experimental results showed that the performance of the system is very satisfactory.

Based on the above work, we established an Artificial Intelligence medical system to diagnose NSCLC patients efficiently using SVM and ANN. Compared with these works, the system we built not only provides the basic functions of diagnosing the NSCLC patients, but also includes an evaluation model for treatment plan.

III. SYSTEM MODEL DESIGN

A. FRAMEWORK OF THE SYSTEM

The medical decision system is designed to assist doctors in making a fast and accurate diagnosis of patients with NSCLC. It contains four functions, diagnosing the patients, staging cancer, recommending the treatment plan, and evaluating the treatment plan. Diagnosing the patients refers to taking three tumor markers as input, using SVM to classify the patient's condition, and telling whether the tumor is benign or malignant. Staging cancer means determining the clinical stage (I, II, III, or IV) for patients with malignant NSCLC using an ensemble model that combines Neural Networks and Support Vector Machines. As for the recommendation and evaluation of the treatment plan, we regard it as a regression problem. Based on the former ensemble model, the system uses a secondary learner and output a value to evaluate the malignancy of NSCLC, abbreviated as EM value. This EM value stands for the malignant degree of the patient's tumor. The larger the value, the higher the malignancy. Based on this EM value, a quantified evaluation of the development of NSCLC can be constructed. Doctors can use this EM value as a reference to determine whether the current treatment plan is effective enough. If it is not, another treatment plan will be recommended. In addition, due to concurrency and independence in the operation of the system, the system needs to satisfy multiple patients' simultaneous diagnosis requirements. Considering the increasing amount of training data over time, the system model will be retrained to improve its performance after the amount of data has grown to a certain extent.

B. DESIGN OF THE DECISION MODEL

In the hospitals, when the test results of the three tumor markers (CYFRA21-1, CEA, CA-125) are high, the condition of the patient will be worse, and the survival rate will be considerably low, so it is on the contrary. In this medical system, we select three major tumor markers (CYFRA21-1, CEA, CA-125) of NSCLC as the input features and SVM [19]–[21] as the classifier in order to tell whether the tumor is benign or malignant [22]–[24]. For the function of staging the cancer, treatment recommendation, and evaluation, we design an ensemble model that combines DAG (Directed Acyclic Graph)-SVM [25], [26] and Artificial Neural Networks with different structures. SVM is a binary classifier. In order to extend it to a multiclass problem, we choose DAG -SVM. The main two kinds of ANN adopted in the system are Multi-Layer Perceptron (MLP) [27]–[29] neural network and Radial Basis Function (RBF) [30]–[32] neural network. Based on the practical experience, we determined the penalty parameter of SVM is 80 by adjusting the parameter. For MLP, it is a model proposed to make up for the shortcomings of Single-layer perceptron network which can only be used for linearly separable classification problems, and its activation function is usually Rectified Linear Unit (ReLU). For RBF, it is a widely used neural network which can avoid falling into local minimum when the training is in progress. And the activation function we used is Radial Basis Function (RBF).

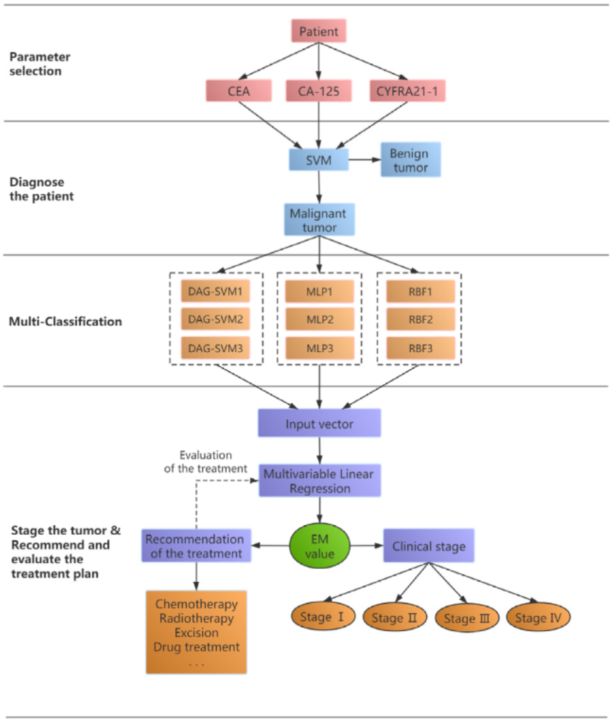

FIGURE 1. **The overall flow of the system.**

The overall flow of the medical system is shown in Figure 1. First, we construct an input vector $\mathbf{x} = (x_{CEA}, x_{CA-125}, x_{CYFRA21-1})$ with three major tumor markers and use SVM to judge whether the tumor is benign or malignant. In the actual diagnosis, when the level of a single tumor marker is high, it does not mean that the tumor is malignant for sure. Excessive alcohol consumption and medication may also cause an increase in level. Here, we take three major tumor markers into consideration for higher accuracy. If the tumor is judged as benign, the system will give recommendations of the following examination and treatment.

If the system determines the tumor is malignant, the ensemble model will be adopted for tumor staging according to the international TNM staging standards [33]. The clinical stage of NSCLC can be divided into four stages: I II III IV, which is actually a four-classification problem. Here we use an ensemble model that combines SVM and Neural Network to complete. Each group forms a DAG structure and outputs a four-dimensional one-hot vector after the sample has been processed from the root to the leaf of the DAG. In order to improve performance, we use DAG-SVM groups with different kernel functions at the same time. Each group uses different kernel functions, and the SVM models in the same group use the same kernel function. Three kernel functions, linear kernel, polynomial kernel and Gaussian kernel, are adopted in this system.

In the course of training, each SVM in the group is considered as one unit and different structures of MLP neural networks and RBF neural networks are used in the system. Each SVM in the system is trained and tuned until the error is below threshold ε. Since three input features are used and the malignant tumor is divided into four classes, the input layer and output layer of MLP and RBF have three units and four units respectively. The numbers of hidden layers and hidden units of them are searched so that each of them can have an error lower than ε. The final structure of MLP neural networks are 3-9-7-4, 3-10-7-5-4, 3-7-5-4 (The number represents the unit number of each layer). In RBF neural networks' structures, they are 3-10-4, 3-14-4, 3-16-4. Then the center c_i of each hidden unit will be determined after clustering the sample data by the k-means algorithm.

Finally, we combine the outputs of the SVM groups, MLP networks and RBF networks as the input vector of the secondary learner. Since each base learner has a 4-dimension output, the dimension of the input vector in ensemble model is 36 (9 base learners by 4-dimensional output). After analyzing the level of tumor markers, we could find that the tumor marker levels are close to the normal range for patients in stage I and benign tumors. For patients in stage III and IV, the levels are far away from the normal range. Therefore, we can suppose that the level of tumor markers of NSCLC conforms to the exponential law. In fact, tumor is hard to find at the early stage, but at the middle and late stages, tumor will spread savagely throughout the body of the patient, which is the main cause of the sudden increase of tumor marker levels. Meanwhile, in order to improve the model's robustness to normal people and benign tumor cases, the output value of the exponential linear regression model does not start from 1. Supervising numerical labels, 3,4,5,6 are added manually for the input samples of patients in stage I II III IV respectively through a function in Algorithm 1. The advantage of setting the start number from 3 is that the model will be able to distinguish benign tumors from malignant tumors. If the development of malignant tumors is controlled by effective treatment and return into normal value, the model can still seize its tumors marker level instead of just outputting 0. Ultimately, the system will output the evaluation value of NSCLC's malignancy (EM value). The procedure of integrating the results of base learners is shown in Algorithm 1.

Based on the above models, the system will judge the stage of malignant tumors according to the EM value and give the appropriate treatment recommendation. In order to implement the recommendation function, we constructed a database of NSCLC patients' EM values. After a patient's EM value is obtained, the system will first compare this value with other values in the database and then search for the similar cases. Finally the recommended treatment plan will be obtained based on the treatment plan of the similar cases.

Algorithm 1 The Training Process of Secondary Learner

```
Input:
Data set: $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, $y_i \in \{I, II, III, IV\}$
        /*$\mathbf{x}_i$ is the input data and $y_i$ is the class label. */
Classifier:
$C = \{DAG-SVM_1, DAG-SVM_2, DAG-SVM_3, MLP_1, MLP_2, MLP_3, RBF_1, RBF_2, RBF_3\}$
Output: Second learning algorithm $H(\mathbf{x})$ : $\ln(y_{EM}) = \omega^T \mathbf{x} + b$ where $\boldsymbol{\omega}$ is the weight vector of the base learners and $b$ is the bias vector.
Begin
$D' = \Phi$/*$D'$ is the processed data set for training. */
for $i$ in $D$ do
  for $k$ in $C$ do
    $z_{ik} = C_k(\mathbf{x}_i)$;/*$z_{ik}$ is a four dimensional one-hot vector*/
  end for
    $y_i' = f(y_i)$;
    /* Function $f(y_i)$ changes the class label into a numerical value*/
$D' = D' \cup ((z_{i1}, z_{i1}, \ldots, z_{ik}), y_i')$;
end for
use $D'$ to train $H(\mathbf{x})$;
output $H(\mathbf{x})$
End
```

However, it should be noted that our recommended granularity is relatively large and the treatment we recommend is actually a type of treatment plan including radiotherapy, chemotherapy, excision, etc, rather than a specific treatment plan. Moreover, the system will update the patient's tumor marker levels and output of the new EM value. By comparing the difference between the two EM values, the treatment plan's effectiveness can be detected.

IV. EXPERIMENTAL ANALYSIS

In this article, the data cases come from Xiangya Hospital, Xiangya Second Hospital and Xiangya Third Hospital. The detailed information of the data is shown in Table 1. Here, doctor's device in Table 2 means the medical equipment for checking the physical condition of patients with NSCLC and helping doctors diagnose the patients and rescue the patients, which includes Positron Emission Tomography (PET) machine, Electrical Impedance Tomography (EIT) machine, Extracorporeal Membrane Oxygenation (ECMO) machine, etc.

Table 2 reflects the normal range of three tumor markers of NSCLC. These values are essential reference factors for the accuracy of the model. According to the hospital's real situation, the weights of CYFRA21-1, CEA, CA-125 in the system are 35%, 28%,27%. The weight of other parameters is 10%.

In order to find out the best method for predicting the tumor stage, we conducted a contrast experiment based on the same data of 12,186 patients with NSCLC. Figure 2 shows the performances of the SVM, ANN, Naive Bayes Classifier, CART, Lasso, and Elastic Net, respectively. From the data shown in Figure 2, we found that SVM achieved the best scores, and ANN achieved the second-best scores in predicting accuracy. Since we only introduced three major features into the model, which is a low-dimensionality problem, it is not suitable to be processed by Lasso and Elastic Net. However, SVM performs well on both low-dimensional data and high-dimensional data, and it is also suitable for models

TABLE 1. Data collection and type for NSCLC in three hospitals.

Data type	Number
Patient information	2,789,675 items
Outpatient service	968,545 people
Doctors' device in outpatient	28,554,590 items
Be hospitalized	1,676,899 people
People Diagnosis	1,124,561 items
Electronic medical records	5,287,413 items
Doctors' device in clinical	31,427,790 items
Inspection records	179,712 items
Medical laboratory records	9,483,216 items
Routine inspection records	24,287,612 items
Operation records	393,218 items
Drug records	90,631 items

TABLE 2. Normal range of three tumor markers.

Types of tumor marker	Normal range
CYFRA21-1	0 ~ 1.80 ng/ml
CEA	0 ~ 5.00μg/L
CA-125	0 ~ 35.00g/L

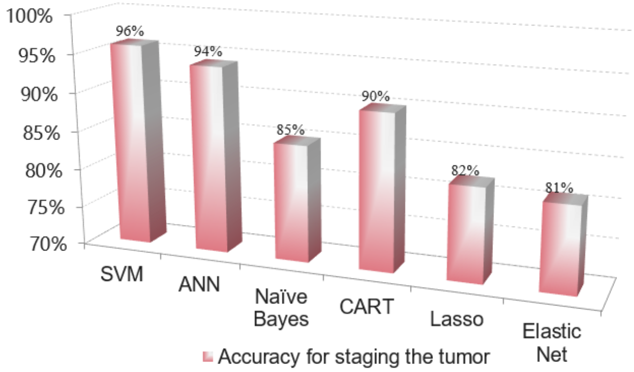

FIGURE 2. Performance of different methods.

whose data size is not so large. Moreover, the model we built also required high accuracy and strong robustness to noise, which is also suitable to be processed by ANN. That is why we adopt SVM and ANN in our model.

The data set is divided into two parts, 80% for the training set and 20% for the test set. First, we select the suitable kernel function and penalty parameter to train SVM. Second, we collect the malignant samples and divide them into four parts according to the clinical stage. For SVM in each group, the malignant samples are represented as $S_{malignant} = \{S_1, S_2, S_3, S_4\}$. For neural networks, the stages of malignant samples are represented as $(1, 0, 0, 0)^T$, $(0, 1, 0, 0)^T$, $(0, 0, 1, 0)^T$, $(0, 0, 0, 1)^T$. Finally, we combine the outputs of the SVM groups and neural networks as the input vector of the exponential linear regression model. Here, we select the mean square loss function as the loss function of this regression model. The function is shown as the following:

$$L(\boldsymbol{\omega}, b) = \frac{1}{m}\sum_{i=1}^{m}(EM_i - EM_i')^2 \tag{1}$$

where EM_i is the evaluation of i-th tumor malignance and EM_i' is the supervising value.

When the training task is finished, we input the malignant samples into the system and calculate the range of EM value for each stage which is shown in Table 3. From these data, we can find that the level of tumor markers of NSCLC conforms to the exponential law.

TABLE 3. EM value of each stage of NSCLC.

Clinical stage of NSCLC	Range of EM	Range of $\ln EM$
Stage I	18 ~ 57	2.9 ~ 4.0
Stage II	58 ~ 119	4.0 ~ 4.8
Stage III	119 ~ 180	4.8 ~ 5.2
Stage IV	> 180	> 5.2

TABLE 4. Contrast of different architectures of the models.

Architecture	Accuracy	Number of misclassified tumors
#1 DAG-SVM+MLP+RBF	89%	1340
#2 2DAG-SVMs+2MLPs+2RBFs	92%	975
#3 3DAG-SVMs+3MLPs+3RBFs	94%	731
#4 4DAG-SVMs+4MLPs+4RBFs	94.6%	658

Table 4 shows the number of the misclassified tumors and the accuracy for staging in four different architectures. As the number of models increases, the accuracy gradually improves. However, starting from Architecture #3, when the number of models increases, there is only a minor improvement for the accuracy. Considering the requirement for high accuracy and stability of the s and the suitable complexity of the algorithm, we adopted Architecture #3 in our system.

Compared with the CAMPAS-P system [11], the TMCA technique [12] and the ANN model used six tumor markers proposed by Feng *et al.* [13] in the Related Work section, all of which adopted the combination of tumor markers to build a medical system for cancer, our medical system can not only diagnose the NSCLC patients with high accuracy, but also give recommendations of treatment plan. Table 5 reflects the differences of theses methods.

TABLE 5. Contrast of different methods using tumor markers.

	Cancer type	Number of tumors markers	Experimental samples	Accuracy
Our method	NSCLC	3	12,186	97.2%
CAMPAS-P [11]	Pancreatic cancer	9	145	100%
TMCA [12]	Primary cancer	>10	307	83%-88%
Feng's method [13]	Lung cancer	6	210	96.9%

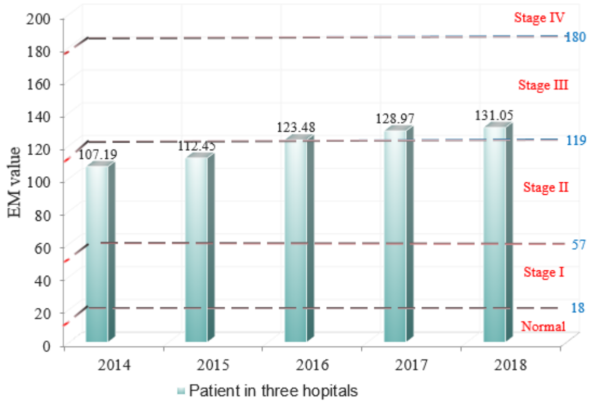

FIGURE 3. Average EM value from 2014 to 2018 in three hospitals.

Figure 3 shows the average EM values of 12,186 patients in three hospitals from 2014 to 2018. From the data, we can find that there is an obvious increase in the number of patients with NSCLC. Moreover, for these five years, the NSCLC cases were mostly in stage III.

In this system, a quantified evaluation of the development of NSCLC is designed. And it's useful for the evaluation and recommendation of the treatment plan by observing the EM value changes. Figure 4 shows the recommended treatment plan and changes in the EM value of a patient with NSCLC. As shown in Figure 4, there is a decrease in the EM value, which proves that the treatment plans recommended by the system were effective, and the malignant tumor of the patient was under control.

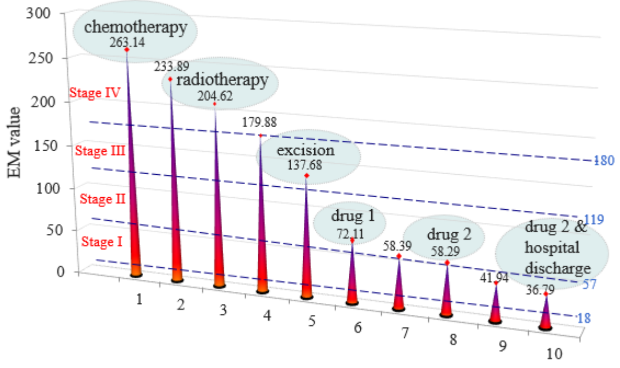

FIGURE 4. A treatment process of a patient with NSCLC.

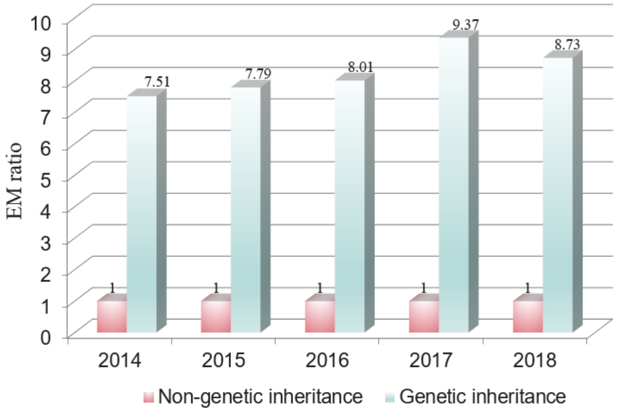

FIGURE 6. Contrast of people with or without genetic inheritance.

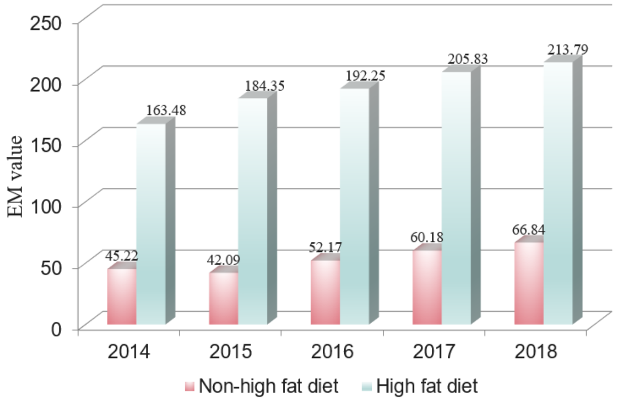

FIGURE 5. Contrast of people with different diet habits.

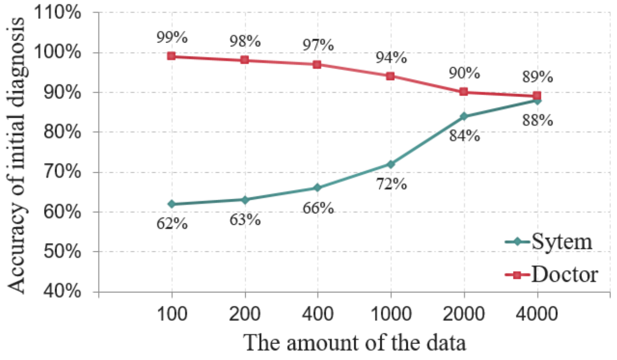

FIGURE 7. Contrast of doctor's accuracy and the system's.

Because our medical system can evaluate the malignancy of the tumor, we can evaluate the influence of a certain factor on NSCLC by observing the value of the control group. Here, we collected some related data and analyzed the influence of patients' diet habits and genetic inheritance on NSCLC. As shown in Figure 5, the diet habit is divided into two parts, a high-fat diet, and none high-fat diet. We can find that the EM value for patients with a high-fat diet is in the range of 163-213, while the value for patients without a high-fat diet is 42-66, which implies the diet habits of patients have an influence on the NSCLC. As shown in Figure 6, the malignancy of patients with genetic inheritance is about 7-8 times more serious than those without genetic inheritance, reflecting the influence of genetic inheritance on NSCLC.

Figure 7 shows the diagnostic accuracy of our medical system and doctors. From the data of 12,186 patients in three hospitals in China, we want to tell whether a patient has NSCLC. For doctors, the diagnosis accuracy rate is very high, nearly 100%, but there is a downward trend of accuracy as the number of patients increases. For our system, when the amount of data is small, the accuracy of the system is 62%-66%, which has a large distance with doctors. However, when the amount of data reaches 4000, the accuracy of the system is almost the same as that of doctors, which implies our medical system is suitable for environments with a large amount of data.

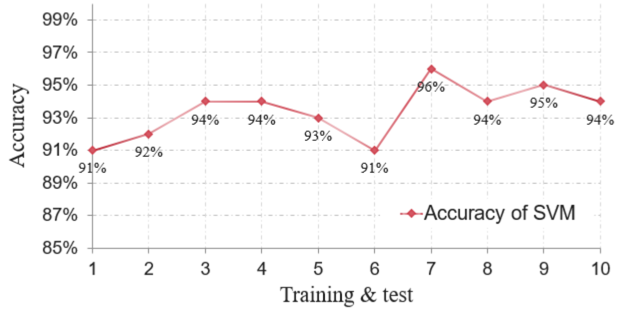

FIGURE 8. Results of 10-fold cross validation.

In order to improve the stability of the SVM accuracy, we performed a 10-fold cross validation by adjusting the data of the training set and the test set based on the data of 12,186 patients. In the 10-fold cross validation, the data was divided into 10 blocks. We randomly chose one block of data as the test set and the others as the training set. Figure 8 shows the results of the 10-fold cross validation. Here, we can calculate the average of SVM accuracy as 93.4%. Figure 9 shows the accuracy of SVM, the number of benign

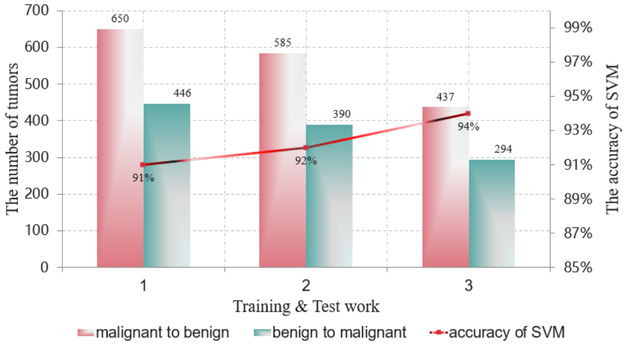

FIGURE 9. Results of training & test work 1-3.

tumors that were misclassified as malignant, and the number of malignant tumors misclassified as benign in the training & test sets 1-3.

V. DISCUSSION

In our proposed study, we design a medical decision system based on Artificial Intelligence, which mainly covers four functions including diagnosing the patients, staging the tumor, recommending the following treatment plan, and evaluating the effect of the current treatment plan. In this medical system, three tumor markers (CYFRA21-1, CEA, CA-125) are selected as the parameters, and classic SVM and ANN models are utilized for classification tasks. Data of 12,186 patients in three hospitals in China served as the sample data in the decision model, which makes the trained model more reliable and practical.

In developing countries, auxiliary medical decision system can not only reduce the waste of medical resources but also extend the patients' survival time by timely diagnosis and evaluation. Simultaneously, doctors' stress is reduced. Thus, the rate of misdiagnosis will decrease.

Based on our study, some benefits are established as follows:

(1) A medical decision system was constructed to assist doctors in diagnosing more efficiently in an environment of a large number of patients.

(2) An evaluation standard for influential factors on NSCLC was proposed. In this study, we tested the effect of diet habits and genetic inheritance. Similarly, we can use this standard to test other factors such as alcohol consumption.

(3) A decision model based on Artificial Intelligence was established for NSCLC, which can be transferred for other cancers such as prostate cancer.

VI. CONCLUSION

In this article, we design a medical decision system based on Artificial Intelligence to assist doctors in diagnosing patients with NSCLC efficiently in developing countries. The system contains four functions, diagnosing the patients, staging the cancer, recommending the treatment plan, and evaluating the treatment plan. Based on data of 12,186 patients in three hospitals, the model of our system has been trained and tested. The result of experiments shows that when the amount of data reaches 4000, the accuracy of the system is close to that of doctors, which proves that our system can be used as an auxiliary diagnostic system to reduce stress on doctors.

However, further studies are needed to improve the accuracy and reliability of the system. Other medical detection information of NSCLC will be introduced into the system.

REFERENCES

[1] 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," *Comput. Methods Programs Biomed.*, vol. 159, pp. 87–101, Jun. 2018.

[2] 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, pp. 138–147, Jun. 2019.

[3] J. L. Mulshine and D. C. Sullivan, "Lung cancer screening," *New England J. Med.*, vol. 352, no. 26, pp. 2714–2720, 2005.

[4] J. Wu, Z. Chen, and M. Zhao, "An efficient data packet iteration and transmission algorithm in opportunistic social networks," *J. Ambient Intell. Humanized Comput.*, vol. 11, pp. 3141–3153, Sep. 2019, doi: 10.1007/s12652-019-01480-2.

[5] I. Kadi, A. Idri, and J. L. Fernandez-Aleman, "Knowledge discovery in cardiology: A systematic literature review," *Int. J. Med. Informat.*, vol. 97, pp. 12–32, Jan. 2017.

[6] S. A. Mostafa, A. Mustapha, M. A. Mohammed, M. S. Ahmad, and M. A. Mahmoud, "A fuzzy logic control in adjustable autonomy of a multi-agent system for an automated elderly movement monitoring application," *Int. J. Med. Informat.*, vol. 112, pp. 173–184, Apr. 2018.

[7] I. Thomas, M. Alam, D. Nyholm, M. Senek, and J. Westin, "Individual dose-response models for levodopa infusion dose optimization," *Int. J. Med. Informat.*, vol. 112, pp. 137–142, Apr. 2018.

[8] 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, Dec. 2017.

[9] C. D. Stylios, V. C. Georgopoulos, G. A. Malandraki, and S. Chouliara, "Fuzzy cognitive map architectures for medical decision support systems," *Appl. Soft Comput.*, vol. 8, no. 3, pp. 1243–1251, Jun. 2008.

[10] T. Muley, T. H. Fetz, H. Dienemann, and H. Hoffmann, "Tumor volume and tumor marker index based on CYFRA 21-1 and CEA are strong prognostic factors in operated early stage NSCLC," *Lung Cancer*, vol. 60, no. 3, pp. 408–415, 2008.

[11] S. Saito, K. Taguchi, and N. Nishimura, "Clinical usefulness of computer-assisted diagnosis using combination assay of tumor markers for pancreatic carcinoma," *Cancer*, vol. 72, no. 2, pp. 381–388, 1993.

[12] T. Kobayashi, "A blood tumor marker combination assay produces high sensitivity and specificity for cancer according to the natural history," *Cancer Med.*, vol. 7, no. 3, pp. 549–556, Mar. 2018.

[13] F. Feng, Y. Wu, Y. Wu, G. Nie, and R. Ni, "The effect of artificial neural network model combined with six tumor markers in auxiliary diagnosis of lung cancer," *J. Med. Syst.*, vol. 36, no. 5, pp. 2973–2980, Oct. 2012.

[14] N. Kureshi, S. S. R. Abidi, and C. Blouin, "A predictive model for personalized therapeutic interventions in non-small cell lung cancer," *IEEE J. Biomed. Health Informat.*, vol. 20, no. 1, pp. 424–431, Jan. 2016.

[15] Y. Pu, M. J. Baad, Y. Jiang, and Y. Chen, "Application of artificial neural network and multiple linear regression models for predicting survival time of patients with non-small cell cancer using multiple prognostic factors including FDG-PET measurements," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2014, pp. 225–230.

[16] M. Tirzīte, M. Bukovskis, and G. Strazda, "Detection of lung cancer in exhaled breath with an electronic nose using support vector machine analysis," *J. Breath Res.*, vol. 11, no. 3, 2017, Art. no. 036009.

[17] 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," *Int. J. Med. Informat.*, vol. 108, pp. 185–195, Dec. 2017.

[18] M. Abdar and V. Makarenkov, "CWV-BANN-SVM ensemble learning classifier for an accurate diagnosis of breast cancer," *Measurement*, vol. 146, pp. 557–570, Nov. 2019.

[19] R. Punmiya and S. Choe, ''Energy theft detection using gradient boosting theft detector with feature engineering-based preprocessing,'' *IEEE Trans. Smart Grid*, vol. 10, no. 2, pp. 2326–2329, Mar. 2019.

[20] V. Cherkassky and Y. Ma, ''Practical selection of SVM parameters and noise estimation for SVM regression,'' *Neural Netw.*, vol. 17, no. 1, pp. 113–126, Jan. 2004.

[21] Y. Chen, X. S. Zhou, and T. S. Huang, ''One-class SVM for learning in image retrieval,'' in *Proc. Int. Conf. Image Process.*, vol. 1, Oct. 2001, pp. 34–37.

[22] D. Moro, D. Villemain, J. P. Vuillez, C. Agnius Delord, and C. Brambilla, ''CEA, CYFRA21-1 and SCC in non-small cell lung cancer,'' *Lung Cancer*, vol. 13, no. 2, pp. 169–176, Oct. 1995.

[23] N. Reinmuth, B. Brandt, M. Semik, W.-P. Kunze, R. Achatzy, H. H. Scheld, P. Broermann, W. E. Berdel, H. N. Macha, and M. Thomas, ''Prognostic impact of Cyfra21-1 and other serum markers in completely resected non-small cell lung cancer,'' *Lung Cancer*, vol. 36, no. 3, pp. 265–270, Jun. 2002.

[24] J. L. Pujol, J. M. Boher, J. Grenier, and X. Quantin, ''Cyfra 21-1, neuron specific enolase and prognosis of non-small cell lung cancer: Prospective study in 621 patients,'' *Lung Cancer*, vol. 31, nos. 2–3, pp. 221–231, 2001.

[25] J. Shen, Y. Jiang, and L. Zou, ''DAG-SVM multi-class classification based on nodes selection optimization,'' *Comput. Eng.*, vol. 41, no. 6, pp. 143–146, 2015.

[26] P. Chen and S. Liu, ''An improved DAG-SVM for multi-class classification,'' in *Proc. 5th Int. Conf. Natural Comput.*, vol. 1, 2009, pp. 460–462.

[27] J. LUO, J. WU, and Y. WU, ''Advanced data delivery strategy based on multiperceived community with iot in social complex networks,'' *Complexity*, vol. 2020, Feb. 2020, Art. no. 3576542, doi: 10.1155/2020/3576542.

[28] J. Wu, Z. Chen, and M. Zhao, ''Community recombination and duplication node traverse algorithm in opportunistic social networks,'' *Peer–Peer Netw. Appl.*, vol. 13, pp. 940–947, Jan. 2020, doi: 10.1007/s12083-019-00833-0.

[29] T. Windeatt, ''Accuracy/Diversity and ensemble MLP classifier design,'' *IEEE Trans. Neural Netw.*, vol. 17, no. 5, pp. 1194–1211, Sep. 2006.

[30] 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,'' *Int. J. Comput. Methods*, vol. 15, no. 01, Feb. 2018, Art. no. 1750082.

[31] M.-L. Zhang and Z.-J. Wang, ''MIMLRBF: RBF neural networks for multi-instance multi-label learning,'' *Neurocomputing*, vol. 72, nos. 16–18, pp. 3951–3956, Oct. 2009.

[32] G.-B. Huang, P. Saratchandran, and N. Sundararajan, ''A generalized growing and pruning RBF (GGAP-RBF) neural network for function approximation,'' *IEEE Trans. Neural Netw.*, vol. 16, no. 1, pp. 57–67, Jan. 2005.

[33] K.-M. Xu and S. K. Krueger, ''Evaluation of cloudiness parameterizations using a cumulus ensemble model,'' *Monthly Weather Rev.*, vol. 119, no. 2, pp. 342–367, Feb. 1991.

HUANZE CHEN is currently pursuing the bachelor's degree in computer science and engineering with Central South University, Changsha, Hunan, China. His research interests include wireless communications and networking, big data research, and medical informatics.

WANGPING XIONG received the B.S. degree from Central South University, China, in 2003, and the M.S. degree from Yunnan University, China, in 2009. He is currently an Associate Professor with the School of Computer Science, Jiangxi University of Traditional Chinese Medicine, China. His research interests include data mining and natural language processing.

JIA WU (Member, IEEE) received the Ph.D. degree in software engineering from Central South University, Changsha, Hunan, China, in 2016. He is currently an Associate Professor with the School of Computer Science and Engineering, Central South University. Since 2010, he has been an Algorithm Engineer at IBM Company, Seoul, South Korea, and Shanghai, China. His research interests include wireless network, big data research, and medical informatics. He is also a Senior Member of the China Computer Federation (CCF) and a member of ACM.

QINGHE ZHUANG is currently pursuing the master's degree with the School of Computer Science and Engineering, Central South University, China. His research interests include wireless networks, big data research, and medical informatics.

GENGHUA YU received the master's degree in information science and engineering from Central South University, Changsha, Hunan, China, in 2017. Her research interests include wireless communications and networking, big data research, wireless networks, and data mining. She was the 2017 Outstanding Graduate of Nanchang University.