Eng Int Syst (2022) 5: 387–398

Engineering
Intelligent Systems

Machine Learning Intelligent Medical Algorithm Based on Computer Vision and Parallel Optimization of Biomedical Information System

Huayong Yang[1,*] **and Xiaoli Lin**[2]

[1] *Department of Information Engineering, Wuhan City College, Wuhan 430083, Hubei, China*
[2] *School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, Hubei, China*

With the development of computer technology, there is a huge conflict between information office and traditional office mode. Using efficient machine learning algorithm for scientific analysis and processing of data is an urgent need in various fields. Although the informatization process in the medical field has been advancing, the related research results are not ideal. This paper aims to process and analyze biomedical information efficiently, provide scientific basis for medical diagnosis, and improve the efficiency of medical diagnosis. Firstly, this paper divides the functions of medical information system from the perspective of computer technology and hospital information management, and designs different functional modules. Then the ER model database is used to build the entity model of each sub database of the system. In order to improve the classification accuracy of high-dimensional data and large-scale data, this paper optimizes the data feature dimension reduction of random forest algorithm, and reduces the dimension of training data set according to the importance of feature variables. The improved random classification algorithm is further optimized in parallel in Apache Spark cloud computing platform, and a parallel random forest algorithm based on parallel Apache Spark is proposed. The proposed distributed parallel random forest algorithm (PRF) is compared with the traditional random forest algorithm RF and DRF in classification accuracy. The highest classification accuracy of PRF algorithm is 0.93 when the number of decision trees is 1500, and the highest classification accuracy is 0.93 when the data size is 1000. When the data volume of data set B is 0.5gb, the execution time of the system is only 18.9s. This shows that the system has shorter execution time and better performance. After using the system, the waiting time of patients in ENT, dermatology, Radiology, pediatrics and oncology increased by 37.93%, 52%, 51.61%, 46.15% and 53.13% respectively, which shows that the use of the system can effectively accelerate the waiting time of patients. 49% of the patients thought that the waiting time was very short, and 72% of the medical staff were very satisfied with the system. This shows that the machine learning intelligent medical algorithm based on computer vision and its biomedical information system are worth promoting.

Keywords: Computer Vision, Machine Learning, Intelligent Medical Algorithm, Biomedical Information System, System Optimization

1. INTRODUCTION

1.1 Background Significance

At present, the traditional machine learning and data mining algorithms cannot directly, efficiently and accurately analyze

*Address for correspondence: Department of Information Engineering, Wuhan City College, Wuhan 430083, Hubei, China, Email: yang_sun1618@163.com

and process a large number of data. In addition, due to the fierce competition in the industry, the demand for fast and efficient business response, most application fields require large-scale data processing with high efficiency, high throughput and high real-time performance [1]. Using parallel computing and cloud computing and other rich computing resources, the design of efficient mechanical learning and data mining technology is also one of the hot topics. With the improvement of China's economic strength, the improvement of national financial strength, the development of hospital health, the demand for medical and health services is also increasing rapidly. People's demand for medical, health and medical services is increasing day by day. Therefore, it is also an urgent problem to speed up the establishment of information, modern and international integrated hospitals.

1.2 Related Work

Many computer vision and medical imaging problems are faced with the problem of learning from large-scale data sets with millions of observations and features. Barbu A proposes a new effective learning scheme, which is based on a criterion and a schedule, and tightens the sparsity constraint by gradually removing variables [2]. His method is generally applicable to the optimization of any differentiable loss function, and has been applied in regression, classification and ranking. Inspired by the actual manual detection process, Bao *Y* proposed a data anomaly detection method based on computer vision and deep learning [3]. Taking the acceleration data of a long-span bridge structural health monitoring system in China as an example, the training process of the method is illustrated, and the effectiveness of the method is verified. Buczak A introduces a centralized literature review of machine learning (ML) and data mining (DM) methods for network analysis supporting intrusion detection [4]. A brief tutorial description of each ML / DM method is provided. This paper discusses the complexity of ML / DM algorithm, discusses the challenges of using ML / DM algorithm in network security, and puts forward some suggestions on when to use the given method.

It is not feasible to rely on the knowledge of doctors to diagnose the symptoms of diseases. Therefore, automatic and intelligent medical diagnosis has become an important work for doctors when dealing with massive and high-dimensional medical databases. Hordri N F proposed a hybrid method based on biogeographical optimization (BBO) and improved MLP learning, and applied it to five kinds of medical diagnosis [5]. He compared the performance of hybrid particle swarm optimization (PSO) and MLP, hybrid genetic algorithm (GA) and MLP, hybrid artificial fish swarm optimization (AFSA) and MLP under the same standard parameters. The specificity, sensitivity, accuracy and precision of each method were evaluated. Medical information is very important, but the computing power of users is limited, so the medical system based on Internet of things is needed to provide users with safe and efficient identity authentication. Park Y H proposes a selective group authentication scheme based on Shamir threshold technology. The selective attribute endows users with the right to choose, making them form a group composed of the things that users choose and visit [6]. Although their research results are insufficient, they all provide a reference for the research of this paper.

1.3 Innovative Points in This Paper

In order to process and analyze biomedical information more efficiently and provide scientific data support for medical diagnosis, this paper studies the parallel optimization of biomedical information system based on machine learning intelligent medical algorithm of computer vision. The innovation of this paper is as follows: (1) from the perspective of computer technology and hospital information management, the function of medical information system is divided, the main function modules include outpatient, inpatient, material and economic, and specific design for different function modules. (2) ER model database is used to build the information entity model of patients and doctors, outpatient medical records, registration forms and examination items. (3) This paper optimizes the random forest algorithm to reduce the dimension of data features, and then proposes a parallel random forest algorithm based on parallel Apache Spark cloud computing platform for further parallel optimization.

2. MACHINE LEARNING INTELLIGENT MEDICAL ALGORITHM BASED ON COMPUTER VISION

2.1 Computer Vision Acquisition and Processing

(1) Image acquisition

Image is the basis of computer vision detection information, and image acquisition is a transformation process from analog video signal to digital signal. Computer vision image acquisition first needs to build a computer vision system. The classic hardware architecture of computer vision system mainly includes light box, light source, camera, image acquisition card and computer [7]. The quality of hardware will affect the results of image acquisition. During the acquisition period, the imaging effect may appear different degrees of distortion and blur. The conventional device of image acquisition is shown in Figure 1.

The light box can avoid the interference of external environment light, provide a stable lighting environment for image acquisition, and ensure the quality of the image. The illumination mode and intensity of the light source will directly affect the quality of the collected image. In order to improve the quality of the collected image and reduce the error rate of the computer system in the subsequent recognition and classification work, we should ensure that the illumination of the target area must be sufficient when designing the light source. Ensure that the target area and the background can be clearly

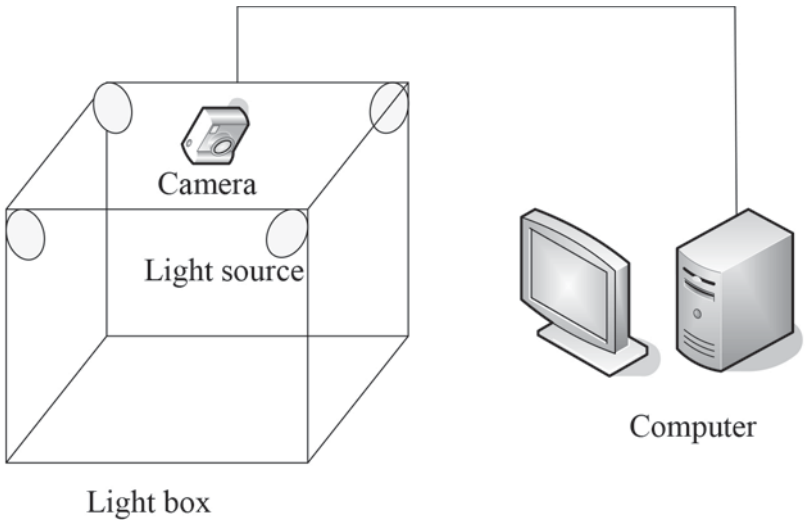

Figure 1 Schematic diagram of image acquisition device.

distinguished, and the image quality is stable. The lighting methods used in the computer vision system generally include foreground light and background light. The foreground light is more uniform and can be flexibly combined, but there are reflections and shadows; the background light imaging has high contrast but lacks surface details [8].

Camera is used for image formation, and photoelectric sensor chip, a key component, can convert optical signal into electrical signal. CMOS camera and CCD camera are commonly used in computer vision system. CCD camera has high sensitivity and signal-to-noise ratio, fast response speed and good response uniformity [9]. Therefore, CCD camera is more suitable for high quality image acquisition scene.

The computer is responsible for the orderly operation of various components and image processing of the quality grading system. The performance of the computer determines the performance of the whole vision system. Therefore, we must meet the basic performance requirements and consider the actual situation to provide hardware support for the scalability of computer vision system.

(2) Image preprocessing

After image acquisition, it needs to be preprocessed to remove the noise caused by environmental factors, enhance the real information of the image, simplify the calculation, and facilitate the subsequent feature extraction. Common image preprocessing methods include image graying, binarization and filtering [10].

The grayscale processing of image can effectively reduce the processing time and speed up the processing speed. The specific methods of graying include component method, maximum value method, average value method and weighted average method. The processing formulas are shown in Formula 1 to Formula 4:

$$f_1(x, y) = R(x, y)\ f_2(x, y) = G(x, y)\ \ f_3(x, y) = B(x, y) \tag{1}$$

$$f(x, y) = \max(R(x, y), G(x, y), B(x, y)) \tag{2}$$

$$f(x, y) = (R(x, y) + G(x, y) + B(x, y))/3 \tag{3}$$

$$f(x, y) = 0.3R(x, y) + 0.59G(x, y) + 0.11B(x, y) \tag{4}$$

Where, $R(x, y)$, $G(x, y)$, $B(x, y)$ represent the R, G and B component values of the point respectively.

Image binarization can highlight useful objects as foreground, distinguish them from background, and minimize the loss of information. Binary processing is based on grayscale processing, which sets the points in the grayscale image to 0 or 255 to display black and white clearly [11]. Binarization generally uses threshold processing, according to the actual needs of the selection of threshold, as a standard and gray image pixel value for comparison, if greater than the threshold, binarization data is 255; if less than the threshold, binarization data is 0. The calculation formula is shown in Formula 5:

$$f'(x, y) = \begin{cases} 255, f(x, y) \geq \omega \\ 0, f(x, y) < \omega \end{cases} \tag{5}$$

Where ω is the set threshold.

Image filtering processing can remove the noise of the image. In the process of image imaging and transportation, it will produce inevitable noise, so it must be filtered [12]. The mean filtering process is shown in Formula 6:

$$S'(x, y) = \frac{1}{N \times N} \sum_{(i,j) \in R(x,y)} S(x, y) \tag{6}$$

Where N is the size of the mask, and the larger the value is, the smoother the image is.

(3) Image feature extraction

Directional gradient histogram (HOG) features have strong robustness to small geometric changes and local contrast changes of the image [13]. When extracting hog features, we need to gray the image first. Then determine the sliding window, divide it into four cells, and calculate the gradient size and direction of each pixel in the cell. The gradient direction is classified and quantized. After counting the features of all the cells, they are concatenated to get the feature descriptor of the sliding window. Finally, the feature descriptors are normalized. The normalization methods include L2 norm normalization, L2 HYS normalization, L1 norm normalization and L1 norm square root normalization. The formulas used in L2 norm normalization and L2 HYS normalization are the same, but they need to be normalized twice. The above normalization formulas are shown in formulas 7 to 9:

$$z \leftarrow z \Big/ \sqrt{\|z\|_2^2 + \alpha^2} \tag{7}$$

$$z \leftarrow z \Big/ (\|z\|_1 + \alpha) \tag{8}$$

$$z \leftarrow \sqrt{z \Big/ (\|z\|_1 + \alpha)} \tag{9}$$

Where z is the feature descriptor of each sliding window, and α is a minimum value to avoid the denominator being 0.

2.2 Intelligent Algorithm Based on Machine Learning

(1) Machine learning

Machine learning mainly imitates people's thinking mode, and its products can be found in all aspects of life. At present, the common machine learning algorithms include classification, clustering, association analysis, etc., which have an unshakable position in the academic circle [14]. With the development and innovation of computer technology, machine learning algorithm has been more widely used in the computer, and the computing speed has gradually improved, which has brought great convenience for daily life and scientific and technological innovation.

Machine learning includes supervised learning, unsupervised learning and semi supervised learning. The data samples in the training dataset of supervised learning have correct classification results. In a large number of training sets, supervised learning can learn a model or function, and then use it to judge new instances. The training set used in the model includes input data and tags. The output of regression analysis model is continuous number, while the output of classification analysis is classification label value [15]. Unsupervised learning only provides data to the computer in the learning process, but does not provide the corresponding value of the data. The computer needs to judge and classify the data to find the feature relationship. Therefore, there is no way to predict the results in unsupervised learning, only output the results. Semi supervised learning needs to construct a classifier, and the input sample features have little labeling information. Because there are a lot of unlabeled data and it is easy to obtain, semi supervised learning algorithm has been gradually developed.

(2) Classification algorithm

Principal component analysis algorithm can reduce the dimension of data space, but it will not damage the data information [16]. We can also get the relationship between variables and use a graph to represent multidimensional data. Regression analysis can be done and variables can be screened. Principal component analysis (PCA) first needs to standardize the data, calculate the correlation coefficient matrix, solve the matrix, determine the number of principal components, and evaluate the generated principal components. The specific operation is shown in formulas 10 to 14:

$$B_{ij} = \frac{x_{ij} - \overline{x}_j}{S_j} \tag{10}$$

$$R = [r_{ij}]_p xp = \frac{B^T B}{n-1} \tag{11}$$

$$|R - \eta I_p| = 0 \tag{12}$$

$$\frac{\sum_{j=1}^{m} \eta_j}{\sum_{j=1}^{p} \eta_j} \geq 0.85 \tag{13}$$

$$U_{ij} = B_i^T a_j^o \tag{14}$$

Where a_j^o is the eigenvector, U_1, U_2, U_3 are the first, second and third principal components.

Random forest classification algorithm is based on decision tree model. Firstly, the training data set is sampled to form N training subsets. Each training subset is trained by a decision tree method, and a decision tree model is constructed [17–18]. N meta decision trees are integrated to form a random forest model. As shown in Formula 15:

$$R(X, \Theta j) = \sum_{i=1}^{k} r_i(x_i, \Theta j) \tag{15}$$

The model of the ith decision tree is $r_i(x_i, \Theta j)$.

Multilayer perceptron adds hidden layer to the neural network composed of single layer neurons, and adopts back propagation method. Its input layer is responsible for transmitting information, and the activation function is the identity function [19]. The output function in the hidden layer is shown in Formula 16:

$$sigmoid(n) = \frac{1}{1 + e^{-n}} \tag{16}$$

The number of output lines of the output layer is shown in Formula 17:

$$soft\max(W_2 x_1 + B_1) \tag{17}$$

According to the above three functions, we can get the MLP function, as shown in Formula 18:

$$f(x) = K(b' + W'(s(b + Wx))) \tag{18}$$

Among them, W', b', W, b are parameters, generally need to use gradient descent algorithm to calculate.

2.3 System Development Environment and Technology

(1) Introduction to Wamp environment

Wamp (Windows + Apache + PHP + MySQL) is a commonly used web framework application system [20]. Apache is a safe and reliable web server, which has the highest market share in the world. It has excellent performance and supports the latest http / 1.1 communication protocol and CGI. Its configuration process is fast, simple and efficient, and can provide users with session tracking. It can also support the virtual host based on IP and domain name, monitor the server in real time, and provide the function of creating personalized server log.

Hypertext preprocessing (PHP) is an open source script for web development, which has the advantages of low threshold and easy learning. Its program design method is object-oriented, drawing on the syntax of C, Java and Perl [21]. PHP can be combined with Apache Web server, with high efficiency. Support cross platform running environment, database access generally through ODBC or its own connection function. Its excellent performance is mainly reflected in high security, cross platform operation, extensive database, fast execution speed and low development cost.

MySQL is a relational database management system, small size, fast, so it is used by many small and medium-sized websites. The excellent performance of MySQL is mainly reflected in its ability to provide multiple programming language APIs, portability, and support multi thread search, multi operating system and multi connection mode [22]. In addition, it also has a visual management tool, which is very simple to operate.

(2) HL7 standard and Middleware Technology

With the reform of medical informatization, the his system of each hospital is more and more huge and complex, and the requirements of data exchange between hospitals are more and more. However, due to the lack of unified standards, the information between hospitals can not be exchanged. In the standard of electronic exchange, health level 7 (HL7) is a mature and the most widely used, based on the internationally recognized medical electronic information exchange [23]. It can develop and develop the transmission protocol and benchmark of hospital information data, standardize the clinical medicine and information management mode, and improve the data exchange and data sharing degree of hospital information system.

HL7 defines the interface standard format according to the structure level of his interface, and can use various current coding standards. Therefore, HL7 is suitable for multiple operating systems and hardware environments, and can also exchange files and data among multiple application systems. HL7 can standardize the format of clinical medicine and management information, reduce the interconnection cost of hospital systems, and improve the level of information sharing between hospital information systems.

Middleware is an independent system software or service program between the operating system and the application, and distributed applications use the software to share resources and cooperate with different technologies [24]. Combined with the middleware technology of HL7, as shown in Figure 2, it can realize convenient information sharing and exchange between internal and external information systems of the hospital, reduce the coupling between various systems, improve the independence of each stage, and have the ability of high productivity, security and good expansibility of the system.

(3) C / S structure and B / S structure

Network software system includes client / server (C/s) structure and browser / server (B/s) structure [25]. As for the server of C / S structure, there are many cases of configuring high-performance PC, workstation, minicomputer and large-scale database system. By exchanging tasks between client and server, the hardware environment is fully utilized. The disadvantage is that we need to write different versions according to different operating systems, so the comprehensive cost is very high.

The client of B / S structure only needs to install a browser, and the server needs to install the software of database and web server. In the B / S mode, the browser of the client exchanges information through the web server, and the database server processes the interactive data. B / S mode is a three-tier distributed system, namely browser web server database server. By completing the main logic in the server, the load of the client computer can be reduced, the cost of maintenance and upgrade can be reduced, and the cost of comprehensive development can be reduced.

Therefore, compared with C / S structure, the development of B / S structure is more simple and convenient, with strong sharing and portability. And it has the characteristics of distribution, as long as the network is provided, it can handle the business. The extension function and management and maintenance of B / S structure are very convenient. Adding a web page can expand the function, logging in the background can realize the management and maintenance, and the client does not need to update.

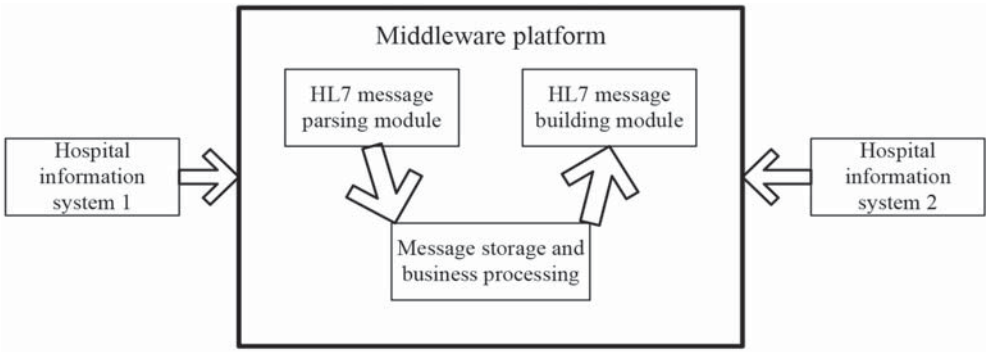

Figure 2 Schematic structure diagram of intermediate parts.

3. PARALLEL OPTIMIZATION EXPERIMENT OF BIOMEDICAL INFORMATION SYSTEM

3.1 Establishment of Medical Information System

(1) Overall design of the system

Medical information system covers all aspects of the hospital. This paper divides the functions of the system from the perspective of computer technology and hospital information management. The main function modules of medical information system include outpatient, inpatient, material and economic. Patients and hospitals are the main body of this system, patients pay for medical expenses, hospitals provide medical services.

(2) System function module design

The outpatient management system not only meets its own business needs, but also provides data to other modules. From the perspective of the outpatient management system itself, the outpatient management system serves the outpatient diagnosis and treatment department, and the data collected by the outpatient management includes the basic information of patients, such as outpatient medical records, registration information, laboratory test results and hospitalization information. The system provides decision-making data to hospital managers. It mainly includes identity registration, appointment registration and outpatient arrangement.

In the traditional sense, the most important stage of inpatient management is to arrange a series of orderly management measures for all aspects of inpatient treatment. The hospital information system can make statistics and query on the flow of patients in the Department at any time, including the basic information of the admission, discharge and critical patients, as well as the empty beds in the inpatient department. It mainly includes inpatient registration and charges, medical record cataloguing, nurse and doctor workstations, and clinical drugs.

Material management mainly monitors the purchase, storage and use of medical supplies. Health economic management is the income and expenditure of the hospital, mainly responsible for price management, outpatient and inpatient charges, accounting and cost accounting.

(3) Database design

The conceptual design of ER model database can be divided into three steps: first, the local ER model is designed, then the local Er models are integrated into a global ER model, and finally the global ER model is optimized to get the final ER model. The information entity model of patients and doctors includes ID number, name, gender, age, home address and telephone number. The entity model of outpatient medical record includes medical record number, patient name, medical record content, diagnosis time and attending doctor. The registration entity model includes registration number, patient name, registration date and department, attending physician and registration category. The inspection item entity model includes the inspection serial number, doctor, analysis, content, result and charge.

3.2 Distributed Parallel Random Forest Algorithm

In order to improve the classification accuracy of high-dimensional data and large-scale data, this paper optimizes the data feature dimension reduction of random forest algorithm, and reduces the dimension of training data set according to the importance of feature variables. Assuming that the training data set x containing M-dimensional feature variables is given, the information gain rate of each feature variable of the current subset of training data is calculated during the training of each tuple decision tree of random forest model, and sorted according to the descending order rule. Then, after the first k feature variables ($k \leq m$) in the sorting list are selected as the main variables, then $(m - k)$ features are randomly selected from the remaining $(m - k)$ feature variables to form the set of M feature variables of the training object. Therefore, the dimension of the feature variable of the training data set is reduced from M to m.

In order to improve the performance of random forest algorithm, this section will effectively solve the imbalance of algorithm expenditure and work types in the process of design

Table 1 Data sets in machine learning database.

Data set	Number of samples	Number of features	Classification	Original data size	Expand data size
A	150000	100000	15	25.2GB	2.0TB
B	245000	320000	5	2.3GB	1.0TB
C	5000000	200000	20	26.8GB	1.5TB
D	180000	150000	12	16.5GB	1.8TB

Table 2 Classification accuracy of different algorithms.

Number of decision trees	PRF	RF	DRF
10	0.85	0.78	0.79
50	0.9	0.82	0.84
200	0.91	0.81	0.84
500	0.9	0.82	0.85
1000	0.92	0.83	0.85
1500	0.93	0.83	0.86
2000	0.93	0.83	0.86

and implementation. The improved random classification algorithm is further optimized in parallel in Apache Spark cloud computing platform, and a parallel random forest algorithm based on parallel Apache Spark is proposed. From the data parallel and task parallel point of view. In data parallelization, vertical data partition and data reuse are used to reduce the cost of data communication between distributed computing nodes. In the task parallelization, the dual parallel training method is used to train at the decision tree level and the node level of each tree.

4. DISCUSSION ON INTELLIGENT MEDICAL ALGORITHM AND BIOMEDICAL INFORMATION SYSTEM

4.1 Classification Accuracy

(1) Comparison of classification accuracy of different random forest sizes

The proposed distributed parallel random forest algorithm (PRF) is compared with the traditional random forest algorithm RF and DRF in classification accuracy. The relevant information of the data set is shown in Table 1.

The decision trees of different scales are constructed to analyze the classification accuracy of each algorithm under different random forest scales. The results are as follows:

As shown in Table 2, when the number of decision trees is 10, the classification accuracy of the three algorithms are the lowest, and the classification accuracy of PRF algorithm, RF algorithm and DRF algorithm are 0.85, 0.78 and 0.79 respectively. When the number of decision trees is 2000, the classification accuracy of the three algorithms is the highest. The classification accuracy of PRF algorithm, RF algorithm and DRF algorithm is 0.93, 0.83 and 0.86 respectively. This shows that the classification accuracy of PRF algorithm is the highest.

As shown in Figure 3, regardless of the number of decision trees, the PRF algorithm proposed in this paper has the highest classification accuracy. PRF algorithm achieves the highest classification accuracy of 0.93 when the number of decision trees is 1500; RF algorithm achieves the highest classification accuracy of 0.83 when the number of decision trees is 1000; DRF algorithm achieves the highest classification accuracy of 0.86 when the number of decision trees is 1500.

(2) Comparison of classification accuracy of different data sizes

The classification accuracy of PRF algorithm, RF algorithm and DRF algorithm is compared under different data scales.

As shown in Table 3, when the data size is 200000, the classification accuracy rates of the three algorithms are the lowest, and the classification accuracy rates of PRF algorithm, RF algorithm and DRF algorithm are 0.9, 0.78 and 0.81 respectively. When the number of decision trees is 1000, the classification accuracy of the three algorithms is the highest. The classification accuracy of PRF algorithm, RF algorithm and DRF algorithm is 0.93, 0.83 and 0.86 respectively. This shows that the classification accuracy of PRF algorithm is the highest.

As shown in Figure 4, the PRF algorithm proposed in this paper has the highest classification accuracy regardless of the data size. PRF algorithm achieves the highest classification accuracy of 0.93 when the data scale is 1000; RF algorithm achieves the highest classification accuracy of 0.83 when the data scale is 1000; DRF algorithm also achieves the highest classification accuracy of 0.86 when the data scale is 1000. This shows that the smaller the data size, the higher the classification accuracy of the algorithm.

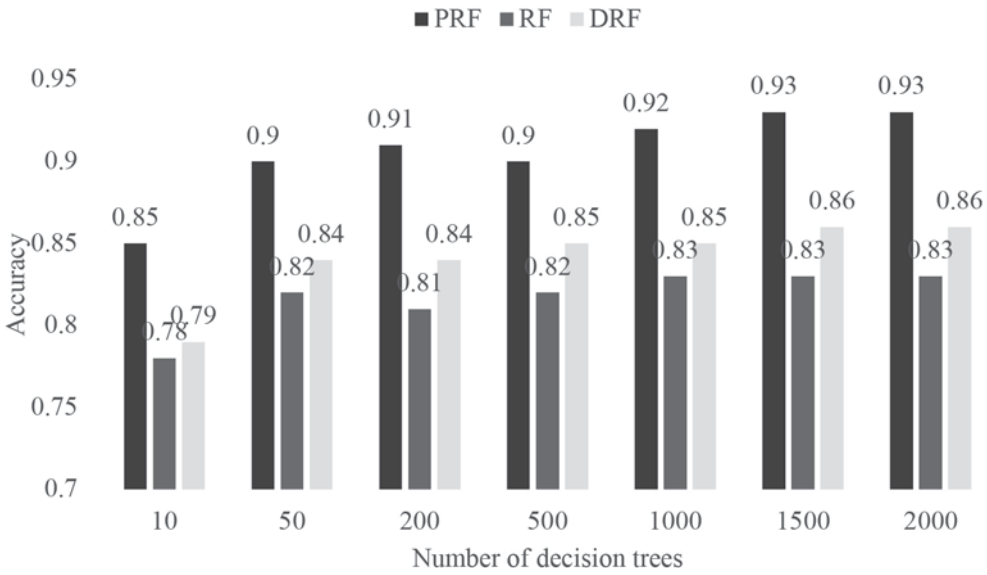

Figure 3 Comparison of classification accuracy under different random forest scales.

Table 3 Classification accuracy of algorithms with different data sizes.

Sample size	PRF	RF	DRF
1000	0.93	0.83	0.86
5000	0.91	0.82	0.85
10000	0.9	0.8	0.85
50000	0.92	0.81	0.83
100000	0.9	0.79	0.82
150000	0.91	0.79	0.82
200000	0.9	0.78	0.81

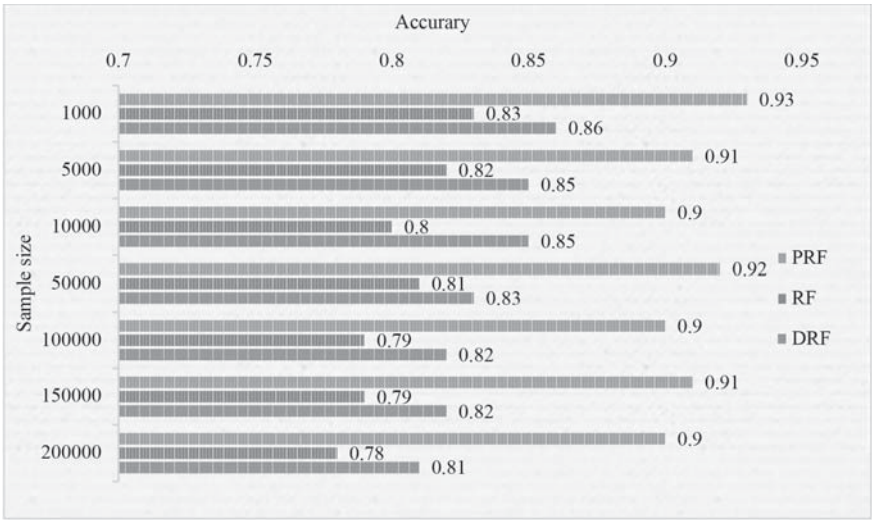

Figure 4 Comparison of classification accuracy under different data scales.

4.2 System Performance Evaluation

The performance of the medical information system is evaluated, and the execution time of the system on a, B, C and D data sets is analyzed. The results are as follows:

As shown in Table 4, when the amount of data is less than 1GB, the average execution time of the system on four data sets does not exceed 23S. When the amount of data is 10GB, the average execution time of the system is 26.635s; when the amount of data is 50GB, the average execution time of the system is 35.375s; when the amount of data is 100GB, the average execution time of the system is 52.975s; when the amount of data is 1000GB, the average execution time of the system is 83.725s.

As shown in Figure 5, the execution time of the system is inconsistent under different data sets of different sizes.

Table 4 System execution time on different data sets (s).

Data size (GB)	A	B	C	D	Average
0.5	23.5	18.9	20.7	21.5	21.15
1	25.1	20.8	23.6	22.1	22.9
10	34.7	22.3	26.1	23.4	26.625
50	51.2	32.2	28.2	29.9	35.375
100	78.3	49.7	42.5	41.4	52.975
500	85.6	70.5	51.2	56.5	65.95
1000	92.3	85.3	78.5	78.8	83.725

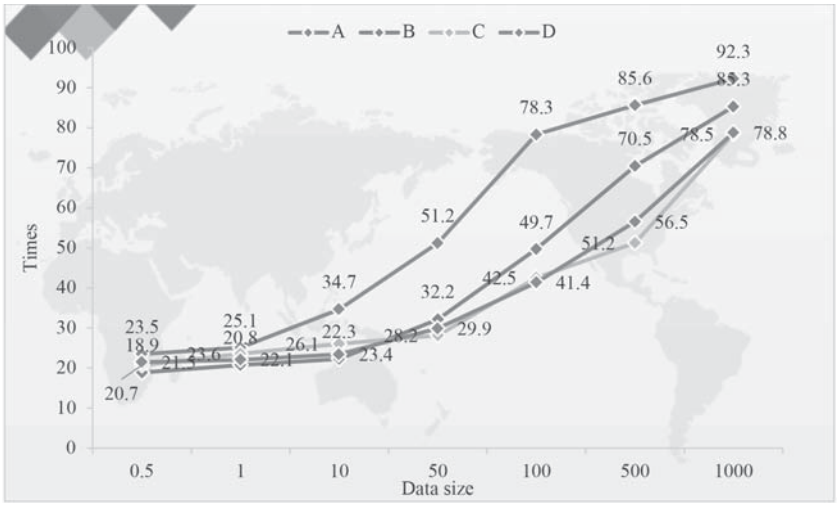

Figure 5 Execution time of the system on different data sets.

Table 5 Average waiting time of patients (min).

Department	Before using the system	After using the system
E.N.T. department	29	18
Dermatology department	25	12
Radiology department	31	15
Pediatrics department	39	21
Oncology department	32	15

No matter which data set, with the increase of the amount of data, the execution time of the system is getting longer and longer. The longest execution time of the system is on dataset a, which takes 92.3s when the amount of data is 1000gb. The shortest execution time is that the system takes 18.9s when the data volume of dataset B is 0.5gb. This shows that the execution time of the system is shorter, the performance is better, and there is room for improvement.

4.3 Application Effect of the System

(1) Waiting time and satisfaction of patients

The system was applied to the actual work of a hospital, and the average waiting time of each patient in five departments of the hospital before and after the system was used was compared (the waiting time here refers to the time for each patient to wait for the last patient to see a doctor), which were E.N.T. department, Dermatology, Radiology, Pediatrics and Oncology.

As shown in Table 5, the waiting time of patients in five departments is greatly shortened after using the system. Before using the system, the waiting time of patients in E.N.T., Dermatology, Radiology, Pediatrics and Oncology was 29 min, 25 min, 31 min, 39 min and 32 min respectively; after using the system, the waiting time of patients was shortened to 18 min, 12 min, 15 min, 21 min and 15 min.

As shown in Figure 6, after using the system, the waiting time of patients in ENT department increased by 37.93%; after using the system, the waiting time of patients in dermatology department increased by 52%; after using the system, the waiting time of patients in radiology department increased by 51.61%; after using the system, the waiting time of pediatric patients increased by 46.15%; after using the system, the waiting time of oncology patients increased by 53.13%. This

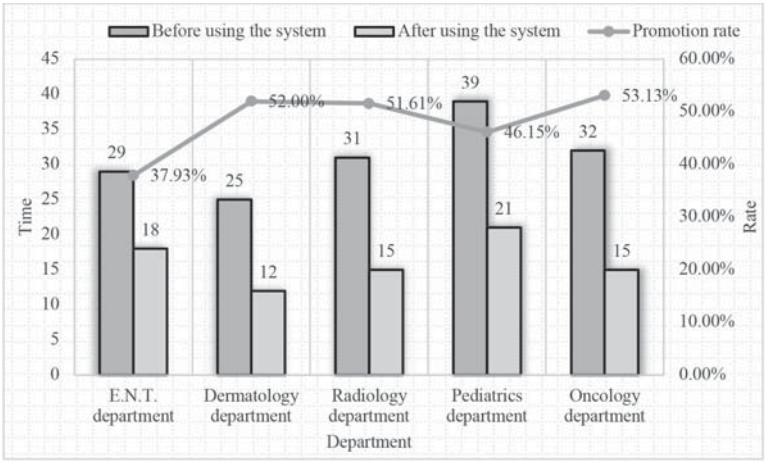

Figure 6 Improvement rate of waiting time of patients.

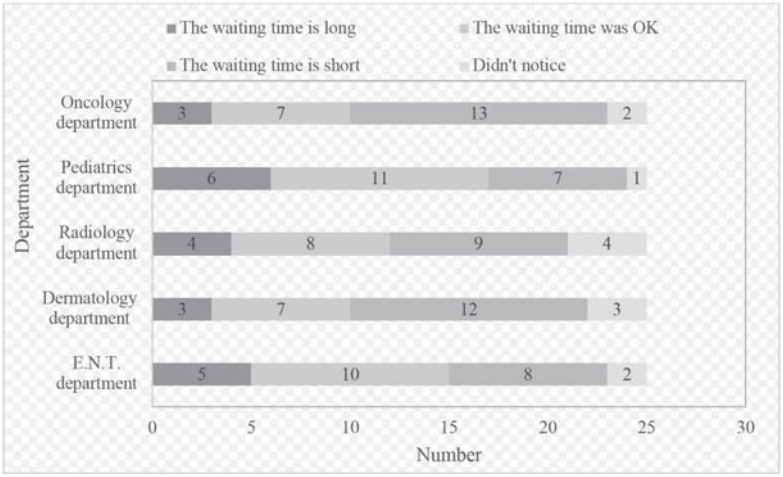

Figure 7 Patients' views on waiting time.

shows that the use of the system can effectively speed up the waiting time of patients.

25 patients in 5 departments were randomly investigated, and their views on waiting time were inquired, which were divided into long waiting time, good waiting time, short waiting time and no attention. The results are as follows:

As shown in Figure 7, 5, 3, 4, 6 and 3 people in Otolaryngology, dermatology, Radiology, pediatrics and oncology think the waiting time is very long; 10, 7, 8, 11 and 7 people think the waiting time is OK and acceptable; 8, 12, 9, 7 and 13 people think the waiting time is very short. Overall, 21% of the 100 patients thought that the waiting time was very long, 43% thought that the waiting time was OK and acceptable, 49% thought that the waiting time was very short, and the remaining 12% did not pay attention to their waiting time.

(2) Experience of medical staff

The overall satisfaction can be divided into a to D, which are very satisfied, satisfied, general and dissatisfied.

As shown in Figure 8, there is 15 and 21 medical staff who is very satisfied with the system, accounting for 72% of the total number. There are 10 people who think the system is average, accounting for 20%. There are four people who are not satisfied with the system, think that the system is too dependent on computers, and learning how to use computers is not a simple thing.

5. CONCLUSIONS

Machine learning mainly imitates people's thinking mode, and its products can be found in all aspects of life. At present, the common machine learning algorithms include classification, clustering, association analysis and so on, which have an unshakable position in the academic community. With the development and innovation of computer technology, machine learning algorithm has been more widely used in the computer, and the computing speed has gradually improved, which has brought great convenience for daily life and scientific and technological innovation. Using parallel computing and cloud computing and other rich computing resources, the design of efficient mechanical learning and data mining technology is also one of the hot topics. With the improvement of China's economic strength, the improvement of national financial strength, the development of hospital health, the demand

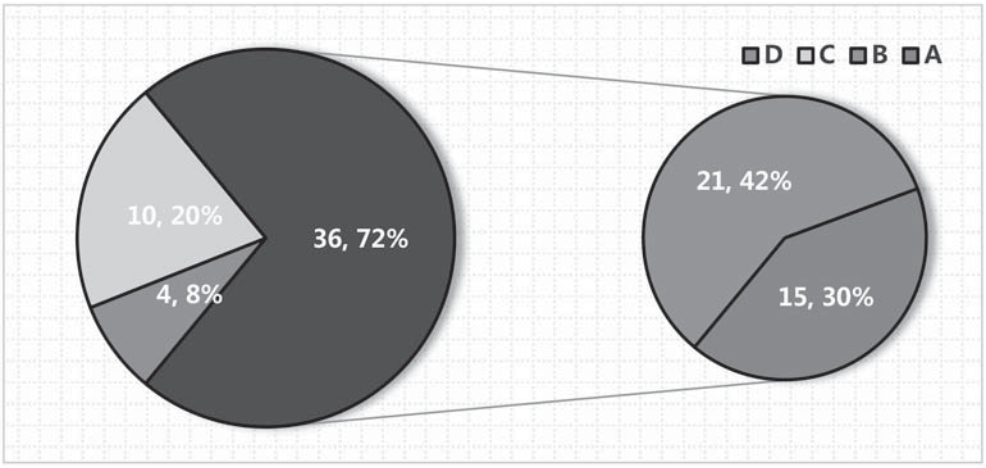

Figure 8 Satisfaction of medical staff to the system.

for medical and health services is also increasing rapidly. People's demand for medical, health and medical services is increasing day by day. Therefore, it is also an urgent problem to speed up the establishment of information, modern and international integrated hospitals.

Medical information system covers all aspects of the hospital. This paper divides the functions of the system from the perspective of computer technology and hospital information management. The main function modules of medical information system include outpatient, inpatient, material and economic. The outpatient management system not only meets its own business needs, but also provides data to other modules. One of the most important stages of inpatient management is to arrange a series of orderly management measures for all aspects of inpatient treatment. Material management mainly monitors the purchase, storage and use of medical supplies. Health economic management is the income and expenditure of the hospital, mainly responsible for price management, outpatient and inpatient charges, accounting and cost accounting.

From the perspective of computer technology and hospital information management, this paper divides the functions of medical information system, and designs different functional modules. Then the ER model database is used to build the entity model of each sub database of the system. This paper optimizes the random forest algorithm to reduce the dimension of data features, and proposes a parallel random forest algorithm based on parallel Apache spark. The experimental results show that the execution time of the medical information system is short and the performance is better. It can effectively speed up the waiting time of patients, and the satisfaction of medical staff is also high.

ACKNOWLEDGEMENTS

This work was supported in part by National Natural Science Foundation of China (No.61972299), by Hubei Province Natural Science Foundation of China (No. 2018CFB526, 2019CFB797), by Philosophy and Social Science Research Project of Education Department of Hubei Province (No. 19G123).

REFERENCES

1. S. Deng, J. Junyong, Z. Lin, et al., Design of a clustered data-driven array processor for computer vision. *High Technology Letters v.26*(04) (2020),78–88.
2. A. Barbu, Y. She, L. Ding, et al., Feature Selection with Annealing for Computer Vision and Big Data Learning. *IEEE Transactions on Pattern Analysis & Machine Intelligence 39*(2) (2017), 272–286.
3. Y. Bao, Z. Tang, H. Li, et al., Computer vision and deep learning-based data anomaly detection method for structural health monitoring. *Structural Health Monitoring 18*(2) (2019), 401–421.
4. A. Buczak, EA. Guven, Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. *IEEE Communications Surveys & Tutorials 18*(2) (2017), 1153–1176.
5. NF. Hordri, SS. Yuhaniz, SM. Shamsuddin, et al., Hybrid Biogeography Based Optimization-Multilayer Perceptron for Application in Intelligent Medical Diagnosis. *Journal of Computational & Theoretical Nanoscience 23*(6) (2017), 5304–5308.
6. YH. Park, A Selective Group Authentication Scheme for IoT-Based Medical Information System. *Journal of Medical Systems 41*(4): (2017), 48.
7. CZ. Dong, O. Celik, FN. Catbas, Marker-free monitoring of the grandstand structures and modal identification using computer vision methods. *Structural health monitoring 18*(5/6) (2019), 1491–1509.
8. A. Taheri-Garavand, S. Fatahi, F. Shahbazi, et al., A nondestructive intelligent approach to real-time evaluation of chicken meat freshness based on computer vision technique. *Journal of Food Process Engineering 42*(4) (2019), e13039.1–e13039.10.
9. Cheng Y P, Li C W, Chen Y C. Apply computer vision in GUI automation for industrial applications. *Mathematical Biosciences and Engineering*, 2019, *16*(6):7526–7545.
10. N. Sengar, MK. Dutta, CM Travieso, Computer vision based technique for identification and quantification of powdery mildew disease in cherry leaves. *Computing 100*(11) (2018), 1–13.
11. BE. Mneymneh, M. Abbas, H. Khoury, Evaluation of computer vision techniques for automated hardhat detection in indoor construction safety applications. *Frontiers of Engineering Management 5*(002) (2018), 227–239.
12. N. Dorbe, A. Jaundalders, R. Kadikis, et al., FCN and LSTM Based Computer Vision System for Recognition of

Vehicle Type, License Plate Number, and Registration Country. *Automatic Control & Computer Sciences 52*(2) (2018), 146–154.
13. LH. Juang, M. Ni, et al. Gender Recognition Based on Computer Vision System. *Intelligent Automation and Soft Computing 24*(2) (2018), 249–255.
14. C. Helma, T. Cramer, S. Kramer, et al., Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds. *J Chem Inf Comput 35*(4) (2018), 1402–1411.
15. Q. Zang, K. Mansouri, AJ. Williams, et al. In Silico Prediction of Physicochemical Properties of Environmental Chemicals Using Molecular Fingerprints and Machine Learning. *Journal of Chemical Information & Modeling 57*(1) (2017), 36–49.
16. Y. Qian, S. Chen, J. Li, et al., A Decision-Making Model Using Machine Learning for Improving Dispatching Efficiency in Chengdu Shuangliu Airport. *Complexity 2020*(8) (2020), 1–16.
17. R. Zhang, GY. Ding, FQ. Zhang, et al., The Application of Intelligent Algorithm and Pulse Coupled Neural Network in Medical Image Process. *Journal of Medical Imaging and Health Informatics 7*(4) (2017), 775–779.
18. F. Xu, H. Lu, The Application of FP-Growth Algorithm Based on Distributed Intelligence in Wisdom Medical Treatment. *International Journal of Pattern Recognition & Artificial Intelligence 31*(4) (2017), 232–237.
19. DX. Wang, Q. Xu, K, Liu, et al., Novel wolf pack optimization algorithm for intelligent medical treatment personalized recommendation. *The Journal of China Universities of Posts and Telecommunications v.25*(06) (2018), 48–61.
20. P. Keikhosrokiani, N. Mustaffa, N. Zakaria, et al., Assessment of a medical information system: the mediating role of use and user satisfaction on the success of human interaction with the mobile healthcare system (iHeart). *Cognition, Technology & Work 22*(2) (2020), 281–305.
21. R. Amin, SH. Islam, P. Gope, et al., Anonymity Preserving and Lightweight Multimedical Server Authentication Protocol for Telecare Medical Information System. *IEEE Journal of Biomedical and Health Informatics 23*(4) (2019), 1749–1759.
22. T. Tan-Hsu, G. Munkhjargal, C. Yung-Fu, et al., Ubiquitous Emergency Medical Service System Based on Wireless Biosensors, Traffic Information, and Wireless Communication Technologies: Development and Evaluation. *Sensors 17*(12) (2017), 202–202.
23. CK. Lo, HC. Chen, PY. Lee, et al., Smart Dynamic Resource Allocation Model for Patient-Driven Mobile Medical Information System Using C4.5 Algorithm. *Journal of Electronic Science and Technology 17*(3) (2019), 231–241.
24. A. Perozziello, T. Gauss, A. Diop, et al., [Medical information system (PMSI) does not adequately identify severe trauma]. *Revue Depidemiologie Et De Sante Publique 66*(1) (2018), 43–43.
25. M. Naghipour, M. Langarizadeh, M. Razzazi, Identification of the requirements for designing medical tourism information system of Iran. *Journal of Education and Health Promotion 8*(1) (2019), 118–118.

Huayong Yang was born in Wuhan, Hubei, P.R. China, in 1977. He received the Master degree from Wuhan University, Hubei, China. He is also an associate professor in Department of Information Engineering, Wuhan City College, Hubei, China. His research interests include computer vision, artificial intelligence, big data analysis and bioinformatics.

Xiaoli Lin was born in Xinyang, Henan, P.R. China, in 1980. She received the M.S. and Ph.D. degree in computer science and technology from Wuhan University of Science and Technology, Hubei, China, in 2007 and 2019, respectively. She is also an associate professor at the School of Computer Science and Technology, Wuhan University of Science and Technology, Hubei, China. Her research interests include knowledge discovery with big data, machine learning, computational biology and bioinformatics.