Auxiliary Diagnosis Method of Chest Pain Based on Machine Learning

Wen Gao, Rong Yu, Zhaolei Yu, Zhuang Ma, and Md Masum

***Abstract*—Chest pain is sudden, its pathological causes are complex and various, fatal or non-fatal so that improving the diagnostic accuracy is extremely important in the emergency system of prehospital and hospitals. Therefore, we propose a method of introducing a decision tree, support vector machine, and KNN algorithm in machine learning into the auxiliary diagnosis of chest pain. First select the algorithm with better performance among decision tree, support vector machine, and KNN algorithm; Then compare the classification performance of the CART algorithm, the support vector machine using the Gaussian kernel function, and the K nearest neighbor algorithm using the Euclidean distance to select the best; Finally, through the analysis of the experimental results, the support vector machine algorithm with Gaussian kernel function is obtained. Its detection time and diagnosis accuracy rate are the best among the three algorithms, which can assist medical staff in the emergency system to carry out targeted chest pain diagnosis.**

***Index Terms*—Chest pain diagnosis, machine learning, decision tree, support vector machine, KNN, classification accuracy, comparison selection.**

I. Introduction

Chest pain is a common problem in the Department of Cardiology occupying the first place in the number of emergency outpatients. The number of patients who go to the hospital to see the emergency department due to chest pain attacks accounts for 20%~30% [1]. Diseases with "chest pain" as the main clinical manifestation are very common, ranging from "acute coronary syndrome" that may endanger life to "intercostal neuralgia" without special treatment. It has a wide variety and complex manifestations. These diseases have caused great distress to clinicians in differential diagnosis, risk stratification, and follow-up treatment [2]. In 2014, the "Chinese Expert Consensus on Standardized Evaluation and Diagnosis of Chest Pain" was compiled by the editorial board of the Chinese Journal of Cardiovascular Disease, which standardized the clinical consensus on chest pain [3]. Some experienced doctors have a high diagnostic rate for chest pain, but the diagnostic efficiency is low. At the same time, chest pain is sudden. Whether in per-hospital treatment or emergency treatment, doctors may face emergencies such as insufficient experience, lack of professional knowledge, emergency treatment, and so on.

With the rise of machine learning theory, more and more methods of machine learning are applied to medical-assisted diagnosis, and good research results have also been achieved. Zhang *et al.* [4] used an improved support vector machine method to construct a new kernel function linearly through experiments, which preserved the optimal performance of each basic kernel function to the greatest extent. It is applied to the auxiliary diagnosis of breast cancer disease. The model has higher classification accuracy, sensitivity, and specificity; Wang *et al*. [5] applied the three algorithms of ID3, Classification and Regression Tree (CART), and AdaBoost in machine learning to the auxiliary diagnosis of melanoma. They developed an accurate and effective recognition method for clinical applications. The new medical diagnosis method for malignant melanoma in vivo not only improves the accuracy of early diagnosis of malignant melanoma and reduces the misdiagnosis rate of benign melanoma, but also provides an objective basis for early clinical detection and diagnosis; Cui *et al*. [6] analyzed the research progress in the application of machine learning to clinical diagnosis of spinal diseases, and made good progress in the segmentation of vertebrae and intervertebral disc, positioning and marking of vertebrae, and clinical auxiliary diagnosis of spinal diseases. At present, machine learning is quite mature in the field of medical assisted diagnosis and can be successfully used in clinical diagnosis. However, there are few applications in the field of chest pain. On the one hand, the causes of chest pain are complex and diverse, and on the other hand, there is a lack of data.

We will be based on UCI's open-source data set on heart disease, including 193 cases related and unrelated to chest pain. Meanwhile, relevant theories and algorithms in machine learning are applied to the auxiliary diagnosis of chest pain. Decision tree, support vector machine (SVM), and k-nearest neighbor algorithm are the classical traditional algorithms in machine learning. In this paper, these three algorithms are applied to the classification of chest pain and implemented by MATLAB, which not only ensures the accuracy, but also increases the diagnosis efficiency, and provides a reference basis for the clinical diagnosis of the emergency diagnosis system.

II. General Diagnosis of Chest Pain

The causes of chest pain are complex and vary in severity, so the diagnosis of chest pain is a difficult challenge for physicians. The diagnosis of chest pain can be divided into three parts: medical history collection, physical examination, and auxiliary examination.

In the part of history information collection, when

Manuscript received June 24, 2022; revised October 21, 2022. The work described in this paper was fully supported by a grant from the Department of Science & Technology of Shandong Province (No.2021TSC1092).

Wen Gao, Rong Yu, Zhuang Ma, and MD MASUM are with the School of Information and Electronic Engineering, Shandong Technology and Busi ness University, China (e-mail: wengao@sdtbu.edu.cn, yurong82@qq.com, mazhuang550043@163.com, md.masum.bd@outlook.com).

Zhaolei Yu is with Yan Tai En Bang Electronic Technology Co., Ltd., China (e-mail: 2606442593@qq.com).

DOI: 10.7763/IJET.2022.V14.1207

receiving patients with chest pain, doctors focus on confirming the nature of chest pain in the posterior sternum, precordial area, and lateral chest. Chest pain is mainly characterized by paroxysmal burning pain, knife cutting pain, and crushing pain. The duration of the pain needs to be recorded in detail, such as continuous pain or intermittent pain for seconds, minutes, and hours, and then ask whether there are some accompanying symptoms such as cough, fever, and shock. Finally, the patient's personal information, such as age, gender, weight, as well as the existence of hypertension, diabetes, coronary heart disease, smoking, and other historical information, was recorded [7].

Physical examination information collection is mainly divided into two aspects, one is the collection of patients' blood pressure (T), respiratory rate (P), heart rate (R), temperature (BP), and other vital characteristics of information, this part can complete real-time collection through non-invasive detection equipment; the other part is to check the appearance characteristics of patients, mainly to confirm whether there are herpes, swelling, skin itching and other symptoms by observation, reconfirm the patient's chest pain position by touch pressure, and confirm whether there are voiced, enlarged, clear and other symptoms in the chest by percussion, the stethoscope are used to confirm whether there is a murmur in the lung and heart, which is the method of watching, listening and asking in traditional Chinese medicine [8].

The clinical auxiliary examination information is that after the patient arrives at the hospital, he needs to complete the examinations such as electrocardiogram, myocardial enzymes, chest X-ray, CT, hematuria and stool routine examination, echocardiography, and other examinations in time [9].

By sorting out the above-mentioned diagnosis ideas, a database for collecting diagnostic information of patients with chest pain can be established. The database contains three parts: medical history collection, physical examination, and auxiliary testing. Each part has corresponding characteristic parameters, which are closely related to the cause and category of chest pain. Assuming that the type of chest pain is $y_1, y_2 ... y_n$ and the characteristic of each part is $x_1, x_2 ... x_n$, then there are: $[y_1, y_2 ... y_n] = f(x_1, x_2 ... x_n)$.

f represents the functional relationship, that is, the classification model. Traditional algorithms in machine learning have strong processing capabilities for classification problems, so the auxiliary diagnosis of chest pain can be realized through machine learning algorithms.

III. Experimental Verification

A. Experimental Data

The data set has 303 examples and a total of 14 attribute. Before making the data set, we need to transform it. The type of chest pain is used as a label, and the feature is used for training that overlaps with the diagnosis of chest pain. 193 data sets were extracted, including 143 samples related to chest pain and 50 samples unrelated to chest pain. Considering the balance of samples, the samples unrelated to chest pain were used twice, that is, the number of samples unrelated to chest pain was 100. At this point, the ratio of sample types is about 1.5:1. The characteristics of the samples are shown in Table I.

TABLE I: Data Relationship and Feature Description

Description of the heart disease dataset						
Age	Sex	ChestPain	RestBP	Slope	Oldpeak	Ca
Chol	Fbs	MaxHR	RestECG	targe	ExAng	
Description of the chest pain dataset						
ChestPain	Age	Sex	RestBP	Slope	Oldpeak	Chol
Fbs	MaxHR	RestECG	targe	ExAng		

B. Verification Method

Using a decision tree, support vector machine, and KNN classification algorithm, and comparing the test time and accuracy of the algorithm, the algorithm with the highest diagnostic efficiency is obtained. The validation method is the five-fold cross-validation method, that is, all data are divided into five parts, four of which are taken as training samples each time, and the remaining one is taken as a test sample. After five experiments are repeated, the average value of the total results is calculated for final analysis and comparison, as shown in Fig. 1.

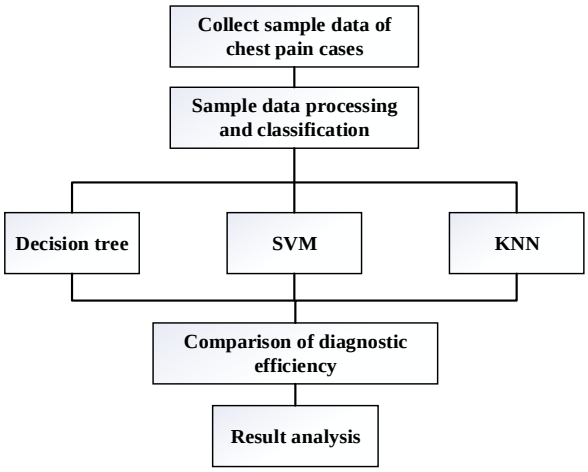

Fig. 1. Experimental flowchart.

C. Decision Tree

A decision tree is one of the most easily understood and commonly used supervised learning classification algorithms in machine learning, and it is also the most intuitive one among various classification algorithms. The Building a decision tree usually uses the information gain of features or other indicators. The category of samples only needs to be judged in turn according to the nodes in the decision tree during classification, so it has the advantages of simplicity and efficiency [10].

The ID3 algorithm of decision tree is suitable for solving the problem of binary classification. C4.5 is improved on the basis of ID3 and can deal with the problem of multiple classification. The classification standard of CART algorithm is different from THAT of ID3 and C4.5, and it uses Gini function as the classification principle to solve both classification and regression problems. A decision tree can effectively deal with missing values without filling them, and avoid data distortion caused by data reduction or improper

filling methods caused by eliminating missing cases. It is suitable for the processing of medical data.

The decision tree uses the above three algorithms to classify the data sets. To solve the problem of poor generalization ability caused by data overfitting, the C4.5 algorithm and cart algorithm is used for pruning respectively. The classification results are shown in Table II.

TABLE II: CLASSIFICATION RESULTS OF DIFFERENT ALGORITHMS

Algorithm name	Classification accuracy(%)	Detection time(s)
ID3	74.4	0.64
C4.5	81.9	0.57
CART	82.3	0.49

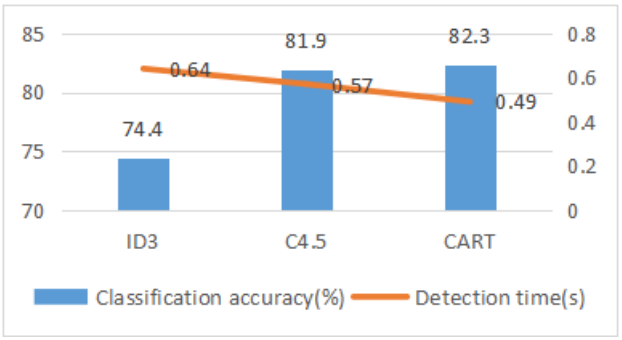

Fig. 2. Comparison chart of classification results.

As can be seen from Fig. 2, C4.5 and the cart algorithm both show a good classification effect, and ID3 has a poor effect. Among them, the cart algorithm has the best classification efficiency.

D. Support Vector Machine

Support Vector Machine (SVM) is a machine learning that method based on Statistical Learning Theory (SET) [11]. Due to its excellent learning performance, it has become a research hotspot in the current machine learning community. It can deal with many problems such as regression (time series analysis) and pattern recognition (classification and discriminant analysis) very successfully and can be extended to prediction and comprehensive evaluation fields and disciplines [12].

SVM is a process to find the optimal plane, which is to map the non-linear separable data to high-dimensional through, the kernel function and find the optimal classification plane in the high-dimensional. Among them, the kernel functions include linear kernel function, polynomial kernel function, Sigmoid kernel function, and Gaussian kernel function. It has its unique advantages in solving high-dimensional and nonlinear problems. At the same time, it's application to small-sample classification and high generalization is worthy of application in medical data mining.

The support vector machine classification algorithm uses the above four kernel functions for training, in which the kernel parameters use the default values of libSVM, and at the same time, the data is normalized pre-processed to improve the classification accuracy of the test. The classification results are shown in Table III.

TABLE III: CLASSIFICATION RESULTS UNDER DIFFERENT KERNEL FUNCTIONS

Kernel function	Classification accuracy(%)	Detection time(s)
Linear kernel function	78.2	0.94
Polynomial kernel function	79.4	0.65
Sigmoid kernel function	82.3	0.71
Gaussian kernel function	86.0	0.49

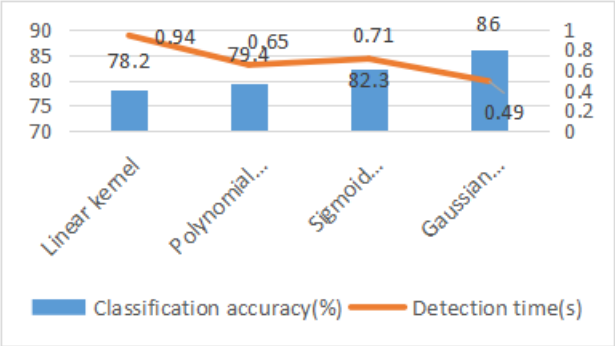

Fig. 3. Comparison chart of classification results.

By comparing the classification accuracy and detection time of support vector machines with different kernel functions, and to verify the performance of the support vector machine classification model, a 5-fold cross-validation method is used to calculate the average accuracy. From Fig. 3, we can intuitively see the difference between the classification accuracy and detection time of the four kernel functions. The Gaussian kernel function has a short detection time and high classification accuracy, so its diagnostic efficiency is the best among the four kernel functions.

E. K-NearestNeighbor Algorithm

The K-Nearest Neighbor (KNN) algorithm is a relatively unique algorithm, a passive classification algorithm, that is, a lazy algorithm that trains the model while testing [13]. Each sample can be represented by its closest k neighbors. By calculating the distance between the data point to be classified and all the data points in the known data set, the first K points with the smallest distance are selected, and the According to the principle of "subject to the majority", the data point is divided into the most frequent category [14].

K-nearest neighbor classification algorithm mainly relies on the nearest neighbor samples rather than the method of discriminating the class domain to determine the sample category to be tested. Therefore, the KNN method is more suitable than other methods for medical data with obvious polymorphism and large class domain crossover. The formula used to calculate the distance is the Euclidean distance and the cosine angle, where the K value is 10, the classification results are shown in Table IV.

TABLE IV: CLASSIFICATION RESULT OF KNN

Category	Accuracy rate(%)	Detection time(s)
Euclidean distance	80.2	0.73
Cosine angle	79.1	0.67

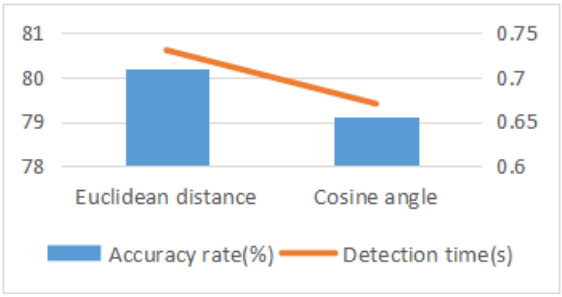

Fig. 4. Comparison chart of classification results.

As can be seen from Fig. 4, Euclidean distance has a slightly higher diagnostic accuracy, but its detection time is longer than the cosine angle. To ensure the accuracy of diagnosis, Euclidean distance is selected as the final comparison object.

IV. Analysis of Experimental Results

To obtain a machine learning algorithm with high diagnostic efficiency, the decision tree uses ID3, C4.5, and CART algorithms for comparison. Support vector machines use different kernel functions, namely linear kernel function, polynomial kernel function, sigmoid kernel function, and Gaussian kernel function for comparison. The KNN algorithm uses different distance calculation formulas to compare when K is a fixed value, and finally compares and selects the algorithm with the best efficiency of the above three algorithms. To study the diagnostic efficiency of various algorithms, the detection time, classification accuracy, and area under the ROC curve of various algorithms are calculated, and 5-fold cross-validation is used to verify the stability of the model. The classification results of the decision tree, support vector machine, and KNN algorithm are shown in Table V~VII.

TABLE V: Classification Result of the Decision Tree

Category	Quantity	Correct rate (%)	1	2	AUC	Detection time	Classification accuracy (%)
1	100	80	80	20	0.88	0.49	82.3
2	143	84	23	120			

TABLE VI: Classification Result of the Support Vector Machine

Category	Quantity	Correct rate(%)	1	2	AUC	Detection time	Classification accuracy (%)
1	100	82%	82	18	0.91	0.49	86.0
2	143	89%	16	127			

TABLE VII: Classification Results of KNN Algorithm

Category	Quantity	Correct rate(%)	1	2	AUC	Detection time	Classification accuracy (%)
1	100	82%	82	18	0.81	0.73	80.2
2	143	79%	30	113			

The number of samples related to chest pain in the data set was 143 and 100 were irrelevant. From the confusion matrix in Tables 3 to 5, it can be concluded that the specificity and sensitivity of the decision tree are relatively stable; the accuracy of the KNN algorithm is the lowest, and its detection time is longer; the specificity, sensitivity, and other indicators of the support vector machine are the highest. The final diagnosis efficiency is determined by comparing the detection time and classification accuracy of the three algorithms. As shown in Fig. 5, it can be seen intuitively that the support vector machine makes a high diagnosis accuracy in a short time.

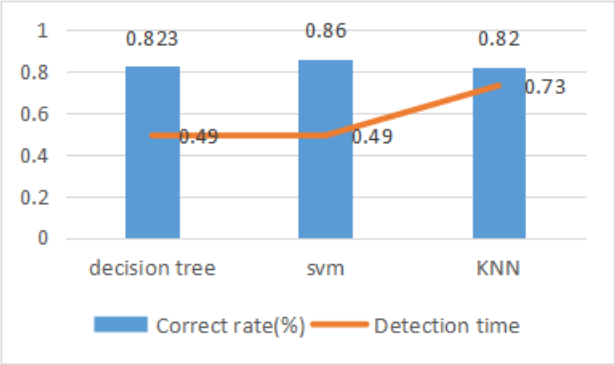

Fig. 5. Comparison chart of classification results.

In summary, the introduction of the machine learning classification algorithm into the chest pain auxiliary diagnosis method proposed in this paper can complete the medical diagnosis on time. It has high diagnostic efficiency and accuracy, and can better meet the outpatient emergency system and pre-hospital emergency Application requirements.

V. Conclusion

In this paper, we introduce a decision tree, support vector machine, and KNN algorithm in machine learning algorithm to solve the deficiency of existing machine learning applications in chest pain-assisted diagnosis. Experiments show that support vector machine has high accuracy and minimum test time, and its diagnosis efficiency is the best. However, the experiment still needs to be improved:

(1) Data set: On the one hand, the number of data sets is small, and on the other hand, the data distribution related to chest pain is too large.

(2) Feature data: In the chest pain center database, the diagnosis of chest pain patients far exceeds the above 11 indicators.

Improving the above two points can not only improve the accuracy but also more accurately predict the type of chest pain. In the hospital emergency system, the diagnosis rate and diagnosis efficiency of chest pain can be improved, and the waiting time of patients can be reduced. It has broad application prospects in various chest pain centers.

Conflict of Interest

We declare that there is no conflict of interest in this paper.

Author Contributions

In this paper, Wen Gao is responsible for conceptual design, Rong Yu is responsible for data analysis and interpretation, Zhaolei Yu is responsible for data collection, Zhuang Ma is responsible for the writing of the paper, and MD MASUM is responsible for polishing the translation of the paper. Finally, all authors agreed to the final version of the paper.

ACKNOWLEDGMENT

Many thanks to the teachers and classmates who helped us in the completion of the paper.

REFERENCES

[1] X. Du and C. S. Ma, "Differential diagnosis and management principles of chest pain," *Chinese Journal of Medicine*, vol. 2003, no. 12, pp. 4-6. DOI: 10.3969/j.issn.1008-1070.2003.12.002.

[2] S. D. Yi, M. Zhou, Y Tian *et al.*, "The enlightenment of the development of guideline of diagnosis and treatment of acute chest pain-from experts consensus to accreditation standards," *China Digital Medicine*, vol. 10, no. 9, pp. 8-10, 2015. DOI: 10.3969/j.issn.1673-7571.2015.09.003.

[3] "Chinese expert consensus on standardized evaluation and diagnosis of chest pain," *Chinese Journal of Cardiology*, vol. 2014, no. 8, pp. 627-632.

[4] Y. L. Zhang, H. B. Shi, W. L. Shang *et al.*, "Improved method for computer-aided diagnosis of breast cancer based on support vector machines," *Application Research of Computers*, vol. 30, no. 8, pp. 2373-2376, 2013. DOI: 10.3969/j.issn.1001-3695.2013.08.033.

[5] T. Wang, N. Zhang, G. R. Hou *et al.*, "Performance comparison of several machine learning methods for computer-aided diagnosis of melanoma," *Application Research of Computers*, vol. 30, no. 6, pp. 1731-1733, 2013. DOI: 10.3969/j.issn.1001-3695.2013.06.034.

[6] Y. X. Cui, Y. Xu, and Q. Fu, "Survey of machine learning applications in the clinical diagnosis of spinal diseases," *Journal of Chinese Computer Systems*, vol. 41, no. 11, pp. 2449-2457, 2020.

[7] Q. Li, "Gender differences in chest pain-related symptoms in patients with acute myocardial infarction," Dissertation, Shandong University, 2018.

[8] X. Y. Ding, M. Wang, and Y. Y. Cai, "Analysis and strategies of skill examination of complete body physical examination in diagnostics," *China Continuing Medical Education*, vol. 13, no. 32, pp. 59-63, 2021.

[9] F. Y. Cai, J. Yang, and J. Wu, "Diagnosis of chest pain in general practice," *Chinese General Practice*, vol. 21, no. 1, pp. 114-118, 2018. DOI: 10.3969/j.issn.1007-9572.2018.01.024.

[10] H. H. Patel and P. Prajapati, "Study and analysis of decision tree based classification algorithms," *International Journal of Computer Sciences and Engineering*, vol. 6, no. 10, pp. 74-78, 2018.

[11] S. Huang, N. Cai, P. P. Pacheco *et al.*, "Applications of support vector machine (SVM) learning in cancer genomics," *Cancer Genomics & Proteomics*, vol. 15, no. 1, pp. 41-51, 2018. DOI: 10.19304/j.cnki.issn1000-7180.2018.05.004.

[12] J. Nalepa and M. Kawulok, "Selecting training sets for support vector machines: A review," *Artificial Intelligence Review*, vol. 52, no. 2, pp. 857-900, 2019.

[13] R. Janani and S Vijayarani, "Text document clustering using spectral clustering algorithm with particle swarm optimization," *Expert Systems with Applications*, vol. 134, pp. 192-200, 2019.

[14] D. Weijie, "Research on classification algorithms in heart disease prediagnosis," Dissertation, Xidian University, 2019.

Gao Wen, PhD, associate professor, master tutor, has long been engaged in teaching and scientific research of electrical engineering and its automation, and has in-depth research in electrical engineering and knowledge engineering, mainly engaged in research and application development of intelligent information processing, knowledge engineering and intelligent software. Lecturing courses such as "Advanced Language Programming" and "Electrical Engineering CAD". He has participated in 2 national 863 projects, 3 national natural funds, 1 EU international project and 1 Beijing project, and hosted 1 Shandong soft science research program project and 4 horizontal R&D projects. At the same time, he has published more than 20 SCI and EI papers, authorized more than 10 national patents, published one monograph, co-edited two textbooks, and won five provincial and ministerial teaching awards.