Machine Learning based Heart Disease Prediction Task

Yijie Zhang*

College of Communication and Information Engineering, Shanghai University, Shanghai, China, 200444

* Corresponding author: 20123909@shu.edu.cn

Abstract. Heart disease is a threat to health condition that has plagued human beings for a long time. The cause of heart disease is complex, and the symptoms are various, which brings many difficulties to the diagnosis and treatment process. With the introduction of machine learning, these algorithms can be used to model the pathogenesis of heart disease and related parameters to complete the initial diagnosis of heart disease. Compared with traditional artificial induction into the causes of heart disease, machine learning-based methods tend to be more efficient and accurate. However, different models may have different performances in this data set since they operate in different ways, so their performances differ from each other. Model accuracy in this task needs to be measured and compared with the same standard. This paper finds that the Random Forest model fits the heart disease prediction task best and has the greatest potential to be optimized. Studying these models which has the most prediction effect on heart disease is valuable for solving this puzzle eventually.

Keywords: Machine learning, heart disease, random forest.

1. Introduction

Heart disease is a non-infectious disease with a high fatality rate which threatens human life and has become one of the prevailing causes of people's death these years [1]. Its onset is slow, its course is long and its pathogenesis is complex. The causes of cardiovascular disease are divers, susceptible to factors like age, sex, degree of hypertension, blood index, etc. The mainstream diagnosing system sometimes faces difficulty in finding and treating such diseases accurately, leading to a shortage in qualified doctors and a slowness of treatment [2]. With data mining technology, more valuable information can be mined from massive medical data collected from previous diagnoses, so that the diagnosis can be more accurate and easier.

Basic models such as KNN, SVM, random forest and CNN have shown their potential in many typical classification and prediction tasks like certain types of heart disease, pneumonia and lung cancer [3]. It is therefore inevitable that these models will also work well in cases where there is a pattern, such as heart disease. The task begins with data processing, where appropriate data will be screened to feed the models. The second step is to import different models to learn from the data and test their accuracies with test data. At last, the model performance will be calculated from the test results so that their value at the task can be measured. In this paper, the performance of several current mainstream machine learning algorithms in heart disease prediction are simulated, tested and compared with each other. This paper provides a general direction for further research in heart disease and gives reference of model selection in such circumstances.

2. Data Processing

2.1. Dataset

The dataset is from UCI, with data collected from 4 regions: Cleveland, Virginia, Hungarian and Switzerland. It contains 14 features: age and sex of the patient; the 3 possible chest pain type of the patient (cp); the blood pressure tested when resting (Trestbps); the amount of cholesterol in the patient's blood (chol); the blood glucose tested when fasting (fbs); the electrocardiogram tested when resting (restecg); the maximum heart rate recorded in a time period (thalach); if the patient suffers

from a angina when exercising or not (exang); the ST inhibition value (oldpeak); the slope of ST segment (slope); the major vessels number (ca); the diagnosed blood disease type (thal); if the patient has a heart disease or not (target). Five lines in dataset are shown in Table 1.

Table 1. The first five unprocessed data

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
32	1	1	95	0	N/A	0	127	0	0.7	1	N/A	N/A	1
34	1	4	115	0	0	N/A	154	0	0.2	1	N/A	7	1
35	1	4	N/A	0	N/A	0	130	1	N/A	N/A	N/A	6	3
36	1	4	110	0	N/A	0	125	1	1	2	N/A	N/A	1
38	0	4	105	0	N/A	0	166	0	2.8	1	N/A	N/A	2

2.2. Data cleaning

The original data set contains a lot of missing data and noise information. For the accuracy of model training, those records with blank values have to be supplemented and the noise data should be removed. To supplement the missing data in the dataset, there are 2 strategies: For discrete data, the blank values (N/A) are filled by the mode of the column; for continuous data, the missing values are filled by the average value of other data of the attribute. A sample will be defined as noise data if there are 5 or more missing values or noise data in it, and the sample will be deleted from the dataset. After these stages of data cleaning, the dataset contains 843 samples.

2.3. Data pre-processing

In order to facilitate the subsequent training, the attributes which have bigger impact on heart disease need to be pointed out. Data types have to be altered to fit these models and the whole dataset have to be divided into training data and testing data.

2.4. Correlation analysis

A heat map shows the correlation between each column of two two-dimensional arrays. To find out the relationship between a single index and the target value, the Pearson correlation coefficient can be used as a standard, which is calculated by the quotient of covariance and standard deviation of two variables [4]. By doing this to all attributes and put the coefficient into a table, the correlation between these attributes is easy to see. The heat map is shown in Figure 1.

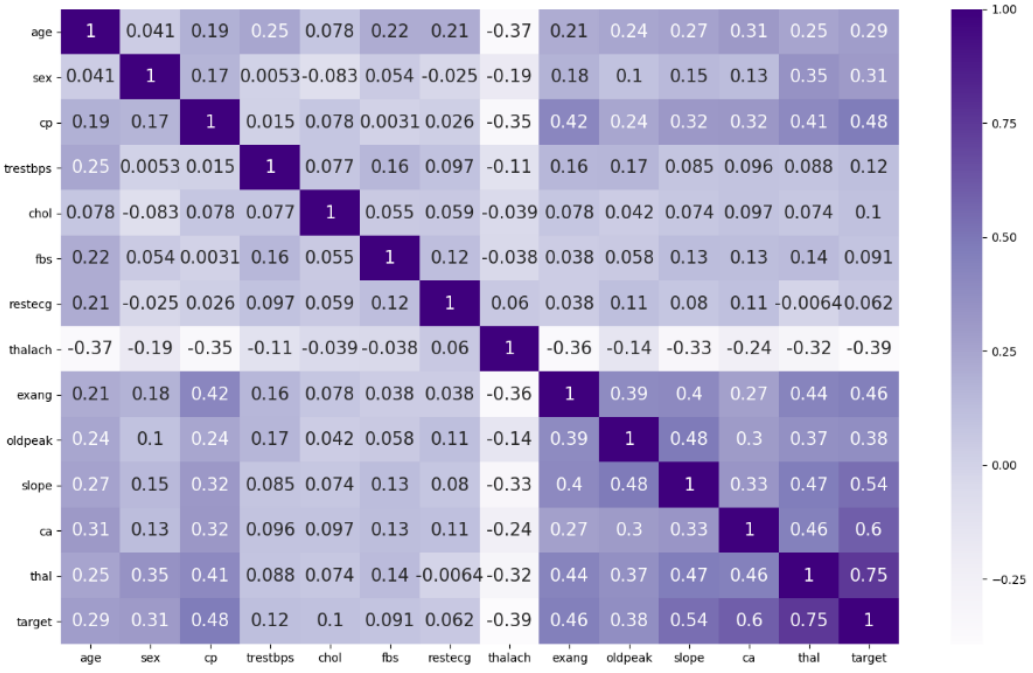

Figure 1. Correlation heat map between all attributes

2.5. Data encoding

The feature categories in the original data are substituted with numerical values, such as 1 for male and 0 for female. During model training, each feature is fitted into continuous data, for example, a number from 0 to 1 is used to represent its gender. As a result, it is difficult to distinguish the specific category in the subsequent steps, which affects the model accuracy.

So, it is therefore possible to encode N feature classes with n-bit code, which is consisted if n register space, and only one of which has the true value in a record. This converts N categories into N binary features, with only one activated at a time. Sex_0 stands for female while Sex_1 stands for male. Cp_1, Cp_2 and Cp_3 represent three different types of chest pain. If fasting blood sugar is higher than 120 mg/dl, fbs_1 is 1, otherwise fbs_0 is 1. Value_0, Value_1 and Vlue_2 stands for the electrocardiographic recorded when resting is in a normal state, having a disorder in ST-T wave or suffering from a left ventricular hypertrophy separately. Exang_1 and exhang_0 indicates whether exercise includes angina. Ca_1 to Ca_9 stands for different numbers of major vessels colored by flourosopy. Thal_3 stands for no blood defect, Thal_6 stands for fixed defect and Thal_7 stands for reversable defect. As is shown in Table 2, the original 8 columns of features will be expanded to 24 columns.

Table 2. Discrete data after one-heat encoding

	Sex_0	Sex_1	Cp_1	Cp_2	Cp_3	…	Ca_3	Ca_9	Thal_3	Thal_6	Thal_7
0	0	1	0	1	0	…	0	0	1	0	0
1	0	1	0	1	0	…	0	0	1	0	0
2	0	1	0	1	0	…	0	0	1	0	0
3	0	1	0	1	0	…	0	0	1	0	0
4	1	0	0	0	0	…	0	0	0	1	0
…	…	…	…	…	…	…	…	…	…	…	…
834	0	1	1	0	0	…	0	0	0	0	1
835	0	1	0	0	0	…	0	0	0	0	1
836	1	0	0	0	1	…	0	0	1	0	0
837	0	1	0	0	0	…	0	0	0	0	1
838	0	1	0	0	0	…	1	0	1	0	0

Distances under some algorithm rules like Euclidean distance and Minkowski distance are crucial to the classification process of many machine learning models. However, whatever is adopted, the distances must be in the same value range in a certain type of model. So data standardization is also an important step of data cleaning. Subtract the feature mean from each data set feature and divide by the feature standard deviation. In this task, the characteristic data is transformed into an interval from -3 to 3.

StandardScaler function is chosen as the standardizing strategy in this task. Firstly, by calculating the mean value and standard deviation of the features in the training set, each feature is independently centered and scaled. Then, the mean and standard deviation are stored and scaled at the same scale on future test sets. After the process, all values share a same range of value like Table 3.

Table 3. Discrete data after one-heat encoding

	age	trestbps	chol	thalash	oldpeak
0	-2.700537	-0.120133	-2.203491	1.824096	-0.795194
1	-2.593120	-0.120133	-0.828432	2.477657	-0.795194
2	-2.593120	-0.404278	-0.179098	1.247375	-0.795194
3	-2.593120	-0.644544	-0.083608	0.862911	-0.795194
4	-2.485703	1.977511	-0.198196	1.247375	-0.795194
…	…	…	…	…	…
834	2.348070	0.194513	-0.427373	-0.982512	1.960180
835	2.348070	1.453100	1.195961	-0.982512	1.041722
836	2.455488	-0.404278	-0.962118	-0.828727	0.215110
837	2.562905	-0.434780	-1.458667	-1.059405	1.041722
838	2.562905	-0.382339	1.081372	0.939804	-0.795194

2.6. Data set partitioning

Considering the limited amount of patient heart disease records collected, 75% of the total data is used for training, while other 25% is used for testing.

3. Models

In this section, models like KNN, SVM, Random Forest and CNN are trained through training sets. At the same time, their effects on the test set are recorded and changes are made to the model parameters based on their effects.

3.1. KNN

K-Nearest Neighbors (KNN) is a simple but classic supervised classification algorithm. During the classification process, the k training samples nearest to the training set will be divided as the same sample group by comparing the distance from themselves to the sample chosen. After rounds of classification, the final prediction is made by all these sample groups.

Euclidean distance is used as distance measure in this experiment because matching problems between objects can be minimized in this way [5]. Euclidean distance measures the absolute distance between two sample points in a multidimensional space. If the dimension of the hyperspace is N, $x_1 = x_{11}, x_{12}, \cdots, x_{1N}$ and $x_2 = x_{21}, x_{22}, \cdots, x_{2N}$ are two samples, then the formula of the Euclidean distance between x_1 and x_2 is

$$L_2(x_1, x_2) = \sqrt{\sum_{n=1}^{N}(x_{1n} - x_{2n})^2} \quad (1)$$

KNN has 4 main steps:

1. Select parameter k;
2. Measure the distance between the sample point to be classified and the given sample point;
3. Choose the nearest k samples;
4. Classify the target sample as the largest number of k samples.

In this case, the majority-voting rule is adopted in the last step, which ensures the model has the highest precision in all situations. However, depending on the quantity of neighbors, the performance of KNN models varies. From Figure 2, a table of neighbor number n and accuracy which indicates how the accuracy score varies with the number of neighbors. The model shows the best accuracy when k value is around 5 to 10. So, k value is 7 during the training process.

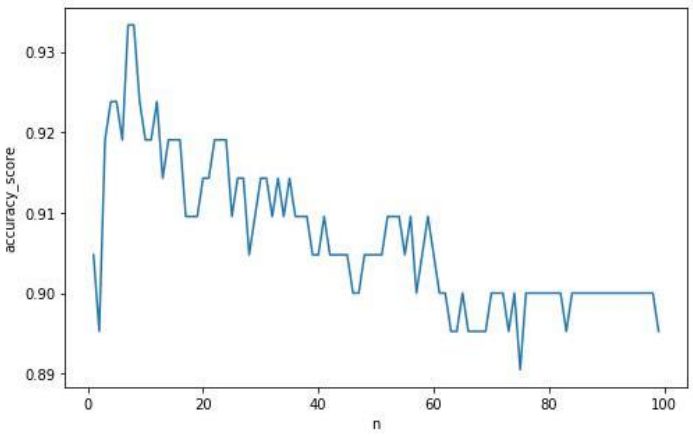

Figure 2. The table of neighbor number n and accuracy

3.2. SVM

Support vector machine (SVM) places the samples in a space of a certain number of dimensions, the coordinate of each point indicates its feature parameters. The SVM model will generate a

hyperplane which distinguishes these types of sample points. The data points at the decision boundary are called support vectors. The hyperplane can be expressed as

$$\boldsymbol{\omega}^T \boldsymbol{x} + b = 0 \tag{2}$$

$\omega = (\omega_1; \omega_2; \cdots; \omega_d)$ is a vector consists of feature parameters of a sample, which shows the direction of the hyperplane. The distance between the plane and the original point is reflected on the value of b. The sum of distances from two different kinds of support vectors to the hyperplane can be expressed by the formula

$$\gamma = \frac{2}{\|\omega\|} \tag{3}$$

The purpose of the model is to maximize the classification interval γ. Then the optimization goal of SVM model can be converted to

$$\max_{w,b} \frac{2}{\|\omega\|} \tag{4}$$

$$\text{s.t.}\, y_i(\boldsymbol{\omega}^T \boldsymbol{x_i} + b) \geq 1, i = 1,2,\cdots,m \tag{5}$$

In practical problems, if the sample cannot be divided by a straight line, the kernel function can be used to raise the dimension. Due to the small size of the training data and small number of features chosen in this task, the Radial Basis Function (RBF) mapping can not only realize the linear division of the original training data into the high-dimensional space, but also have a smaller consumption in calculation [6].

$$\text{k}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}\right) \tag{6}$$

σ is the width of RBF kernel, and$\sigma > 0$

3.3. Random Forest

Random Forest (RF) is inspired by the idea of decision tree. It integrates groups of decision trees and integrates all the prediction results. Each decision tree is acted as a classifier, producing a result to be used in the following step. Then, all results received from the single trees will be gathered and integrate into a final output with the standard of voting criteria [7]. It has 4 main steps:

1 A decision tree is composed of N records extracted from the dataset, and repeated samples are allowed;

2 When the node needs to be split, the decision nodes is selected by chance, but the size of M must be controlled by the amount of parameters. Then select one of the m attributes as the split attribute of the node;

3 Repeat step 2 to continue splitting properties until they can no longer be split;

4 Repeat 1 ~ 3 .

n_estimators are the number of base estimators in the forest, i.e. the number of trees. The n_estimators reflects the performance of a Random Forest model, a higher value of n_estimators means the model has a better precision, but the accuracy tends to be stable when it reaches a certain level. According to Figure 3, the model is generally more accurate when the n_estimators are between 50 and 250. In the subsequent training, the n_estimators are set as 200.

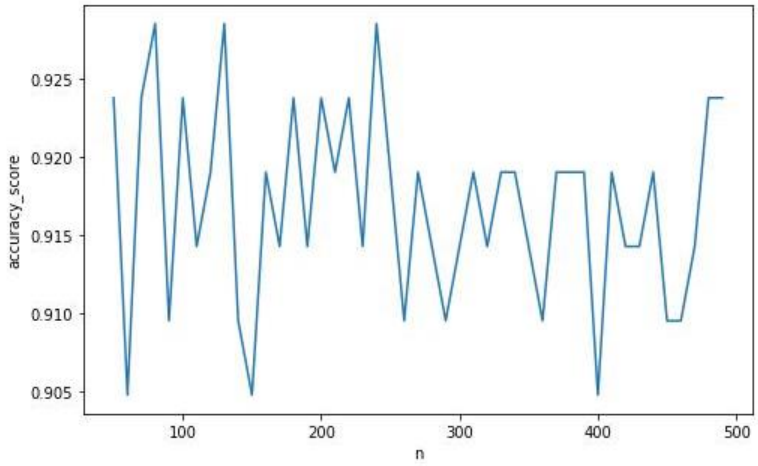

Figure 3. The table of neighbor number n and accuracy

3.4. CNN

Neural Network (NN) is a model that imitates a true neural network structure in a creature's brain. Convolutional neural network (CNN) is one of this series of models which features the complexity and an ability of parallel information processing. It achieves the learning target by adjusting the connecting weight and structure between the learning layers and nodes [8].

The network structure in this task is consisted of 4 fully connected layers, 2dropout layers and an output layer. The four fully connected layers has 256, 128, 64 and 32 dimensions of the output space respectively. And the activation function adopted is ReLU.

$$\text{ReLU(x)} = \begin{cases} x & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{7}$$

When training the standardized data, ReLU function will suppress features less than 0 and strengthen features with large values according to their magnitude. Due to its unilateral inhibition, the function increases the model sparsity. With the unilateral suppression, ReLU evolves fewer calculations, which makes the model more efficient.

In order to avoid overfitting, there are 2 dropout layers after the second and fourth fully connected layer with their dropout rate 0.2 and 0.1 respectively. It means that 20% and 10% of the neurons are discarded by chance in this layer.

The model outputs the one-dimensional output through the output layer. In this layer, the sigmoid function is used as the activation function.

$$\text{Sigmoid} = \frac{1}{1+e^{-x}} \tag{8}$$

The output values of Sigmoid function are limited to 0 to 1, it normalizes the output of each neuron. With its limited output range and stability of the optimization, Sigmoid function can be used in output layer.

Considering the efficiency of model training, batch size is 3 and epoch is 200, which means 3 samples are used in one training session and the training process is repeated 200 times.

One of CNN's advantages is the error back algorithm. The algorithm adjusts the input layer parameters based on the output of each layer through the chain rule [9]. It solves the problem of adjusting the weight of the hidden layer of multi-layer neural network. In this CNN model, the loss function can be expressed as

$$\text{Loss} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(p(y_i)\right) + (1-y_i)\log\left(1-p(y_i)\right)\right] \tag{9}$$

The value of y is 0 or 1and p(y) means the probability of output labels belonging to y. If the predicted p(y_i) approaches 1, then the loss function should approach 0, which enables the model to retain appropriate parameters and modify inappropriate ones.

In the whole training process, the model accuracy rate changes as the number of training times increasing, as shown in the Figure 4.

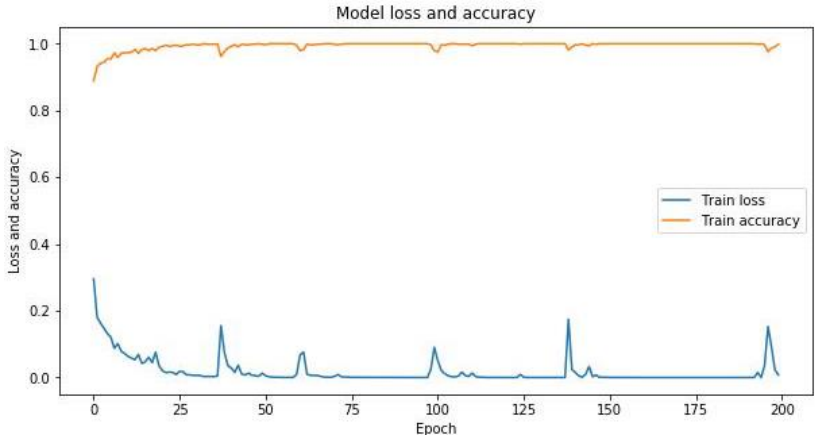

Figure 4. Loss and accuracy of CNN model

4. Results

The confusion matrix can be used to show the training results of supervised learning models. Each column of the confusion matrix represents the prediction category, and each row represents the true attribution category of the data. So it can be divided into 4 parts: True Positive (TP), False Negative (FN), False Positive (FP), True Negative (TN).The confusion matrixes are showed in Table 4.

Table 4. Confusion matrix

(a) KNN

	Prediction Positive	**Prediction Negative**
Reference Positive	89	9
Reference Negative	6	106

(b) SVM

	Prediction Positive	**Prediction Negative**
Reference Positive	75	11
Reference Negative	9	115

(c) Random Forest

	Prediction Positive	**Prediction Negative**
Reference Positive	96	6
Reference Negative	6	102

(d) CNN

	Prediction Positive	**Prediction Negative**
Reference Positive	93	6
Reference Negative	11	100

From the confusion matrix, more advanced classification indexes can be calculated: precision, recall value, F1 score and accuracy score.

Precision value represents the proportion of samples that are truly positive among those simply identified as positive. If the purpose simply focuses on the results, precision will be the main decision parameter. In general, the higher the accuracy, the better the effect of the model is.

$$\text{precision} = \frac{TP}{TP+FP} \tag{10}$$

Recall value represents the portion of positive samples correctly identified by the model to the total number of positive samples. It give a measure of a model from the perspective of input layer. Same as precision, the higher the Recall, the more positive samples will be correctly predicted by the model.

$$\text{recall} = \frac{TP}{TP+FN} \tag{11}$$

F1 score is a measure of classification problems. It is the harmonic average of accuracy rate and recall rate, with the maximum being 1 and the minimum being 0.

$$\text{F1} = \frac{2 \cdot precision \cdot recall}{precision+recall} \tag{12}$$

Though precision, recall and F1 value shows the model performance partially, the overall performance of a model is always decided by accuracy. It represents the accuracy of the total number of samples that the model identifies correctly. Generally, the higher the accuracy of the model, the better the effect of the model is [10].

$$\text{accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{13}$$

Different from precision, recall value and F1 score, accuracy score is the evaluation of the classification effect of all categories, so it can be taken as the most important indicator to measure the performance of the models [11]. The average precision, recall, F1 value and accuracy are showed in Table 5.

Table 5. Average precision, recall, F1 value and accuracy of the four models

	precision	recall	F1 value	Accuracy
KNN	0.93	0.93	0.93	0.9286
SVM	0.93	0.93	0.90	0.9048
Random Forest	0.94	0.94	0.94	0.9429
CNN	0.92	0.92	0.92	0.9190

Figure 5 shows the final accuracy and other evaluation indicators of the four models in a barchart.

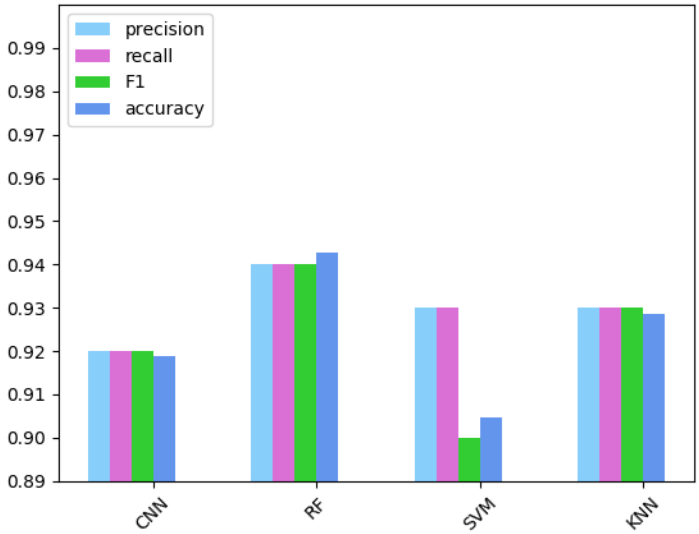

Figure 5. Average precision, recall, F1 value and accuracy of the four models

The precision, recall, F1 value of CNN model are all 0.92, so its final accuracy is 0.9190. Random Forest model has the best performance with the highest precision, recall, F1 value of 0.94 and the accuracy of 0.9429. SVM model has the high precision and recall of 0.93, but it is low in F1 score with only 0.90, so its accuracy is 0.9048 at last. The precision, recall and F1 value of KNN model are all 0.93, so it has a good performance with the final accuracy 0.9286, just below the Random Forest model.

Overall, as can be clearly seen from the bar chart, Random Forest is good in several accuracy indicators, which means that it has the best performance in handling the task of predicting heart disease. KNN model is in the second position while CNN and SVM is less accurate in the prediction task.

5. Conclusion

Models like KNN, SVM, Random Forest and CNN are commonly used in machine learning and are often used for classification, prediction and other purposes. Predictive effects of these models were measured and compared in the case of heart disease prediction. Based on indicators like precision, recall, F1 value and accuracy, all the models have good predictive effect on the training set with the average accuracy over 90%. Among these tested models, the RF model performances better than other models with accuracy score over 94%, which means random forest model has the best application prospect in heart disease prediction. In comparison, KNN and SVM model is low in F1 score, CNN model is less perfect in precision and recall value. When processing high dimensional data such as the heart disease data, random forest does not need to do feature selection, and can conduct parallel computation so that the training speed is faster, unbiased estimation is used in the training process to improve the model generalization. Besides, the structure of a random forest may consist of many similar decision trees, thus masking the real results. In order to enhance the practicality, the size of dataset should be enlarged to ensure that a wide range of etiology is covered though this needs a further cooperation between hospitals and research institutions.

However, much work needs to be done to further improve the model's generality and efficiency. For example, with the idea of integrated learning, training data can be divided into different clusters and use different models for training in order to get a more significant generalization performance.

To sum up, the machine learning approach could greatly aid in the diagnosis of diseases like heart disease. Even several of the most basic and commonly used models tested have shown high accuracy in the case, which proves the great development prospect of machine learning methods in the field of medicine. In the future, integrated models can be used to learn from larger and more extensive disease data to form efficient models for diagnosis of different diseases.

References

[1] Theresa Princy. R, J.Thomas, “Human Heart Disease Prediction System using Data Mining Techniques”, International Conference on Circuit,Power and Computing Technologies [ICCPCT] (2016).

[2] Sun Tiezheng, Yu Zehao. Heart cases based on machine learning classification prediction research [J]. Computer knowledge and technology (2021).

[3] R. G. Franklin and B. Muthukumar, "Survey of Heart Disease Prediction and Identification using Machine Learning Approaches," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS) (2020).

[4] M. J. A. Junaid and R. Kumar, "Data Science And Its Application In Heart Disease Prediction," 2020 International Conference on Intelligent Engineering and Management (ICIEM) (2020).

[5] Zhang Hui, Chen Dandan, Shao Jing et al. Construction, validation and comparison of metabolic syndrome risk prediction model based on KNN algorithm [J/OL]. Chongqing medical: 1-7 (2023).

[6] Timothy M. Amado and Jennifer C. Dela Cruz "Development of Machine Learning-Based Predictive Models for Air Qualit y Monitoring and Characterization" TENCON 2018 - 2018 IEEE Region 10 Conference 2018 (2018).

[7] L. Wei Y. Yang R. M. Nishikawa M.N. Wernick and A. Edwards "Relevance vector machine for automatic detection of clustered microcalcifications" IEEE Trans. Med. Imag. vol. 24 no.10 (2005).

[8] Cousyn, C., Bouchard, K., Gaboury, S., and Bouchard, B. Towards Using Scientific Publications to Automatically Extract Information on Rare Diseases.Mobile Networks and Applications, 1-8 (2019).

[9] S. Kido, Y. Hirano and N. Hashimoto, "Detection and classification of lung abnormalities by use of convolutional neural network (CNN) and regions with CNN features (R-CNN)," 2018 International Workshop on Advanced Image Technology (IWAIT), Chiang Mai, Thailand (2018).

[10] Zhang Xiaoyu, Shen Chao, Lin Chenhao et al. Test and Repair for Machine Learning Model Security [J]. Acta Electronica Sinica (2022).

[11] H. Gaßner et al. "Gait variability as digital biomarker of disease severity in huntingtons disease" Journal of neurology pp. 1-8 2020 (2020).