Hepatitis C Risk Prediction Based on Adaboost

Jingbo Yang *

School of Artificial Intelligence, Xidian University, Xi'an, Shaanxi, 710126, China

* Corresponding Author Email: 21009290058@stu.xidian.edu.cn

Abstract. Hepatitis C is one of the major public health threats. The incidence of liver cirrhosis in 20 years after infection is about 20%, and the annual incidence of hepatocellular carcinoma is 2% - 4%, which is extremely harmful to the health and life of patients. However, people's understanding of hepatitis C is not comprehensive, and only 1 percent of hepatitis C patients worldwide have received effective treatment. At the same time, the early symptoms of hepatitis C are not obvious, and the differences between acute and chronic hepatitis C are large, leading many people to miss the best time for treatment. Therefore, reasonable prediction and classification of hepatitis C at an early stage can provide the most accurate medical guidance for patients and people with related symptoms. Machine learning is widely used in the prediction and classification of diseases in various medical fields, and its maturity has also been widely verified. In this paper, several types of machine learning models represented by decision trees are constructed in Python language to learn and predict the data provided by Ainshams University, and the accuracy rate is 72%. Finally, the data of the data set is analyzed, and relevant suggestions for preventing hepatitis C and in the treatment process are given.

Keywords: Hepatitis C, machine learning, Adaboost.

1. Introduction

Viral hepatitis is one of the major threats to public health in the world. Therefore, in 2016, the World Health Organization proposed to eliminate it by 2030.In July 2021, China issued the "Healthy China 2030 Action White Paper on Eliminating the Threat of Hepatitis C". In September of the same year, China's National Health and Health Commission issued the "Action Plan for Eliminating the Public Health Hazards of Hepatitis C (2021-2030)", and formulated the three-step action goal to eliminate the harm of hepatitis C. The study found that hepatitis C screening is very difficult, and hepatitis C infection has a major feature: the majority of patients have no obvious or atypical symptoms. Some infected people do not pay attention to the early symptoms of hepatitis C, and think it is just common physical discomfort, thus missing the opportunity for early diagnosis [1,2]. Therefore, the biggest difficulty in eliminating hepatitis C is to find out all patients with hepatitis C, and the best way to judge whether they have hepatitis C is to start with the symptoms of patients [3,4]. The existing research is mainly to predict the infection of hepatitis C and determine whether liver fibrosis occurs after infection, but it lacks the judgment of the type of hepatitis C [5,6]. In this paper, according to the specific symptoms and related data of hepatitis C patients, machine learning is used to predict a specific symptom of the patient, and predict whether the patient's hepatitis C is acute or chronic, so as to assist in the early screening of hepatitis C.

2. Method

2.1. Data Set

This data is provided by the medical school of Ain Shams University [7]. The data set provides 1385 valid data of Egyptian patients who have not received HCV dose treatment for 18 months. There are 27 fields in total, including basic information, symptom performance, and DNA and RNA index at the treatment stage. To sum up, this data set can be regarded as a representative sample.

In the data set, the proportion of men and women is roughly 1:1. Figure 1 and Figure 2 show the distribution of patients' age and BMI respectively. It can be seen that the sample distribution of the

dataset is relatively balanced. Also, it reflects the universality of data set sampling and can be used as a sample for research.

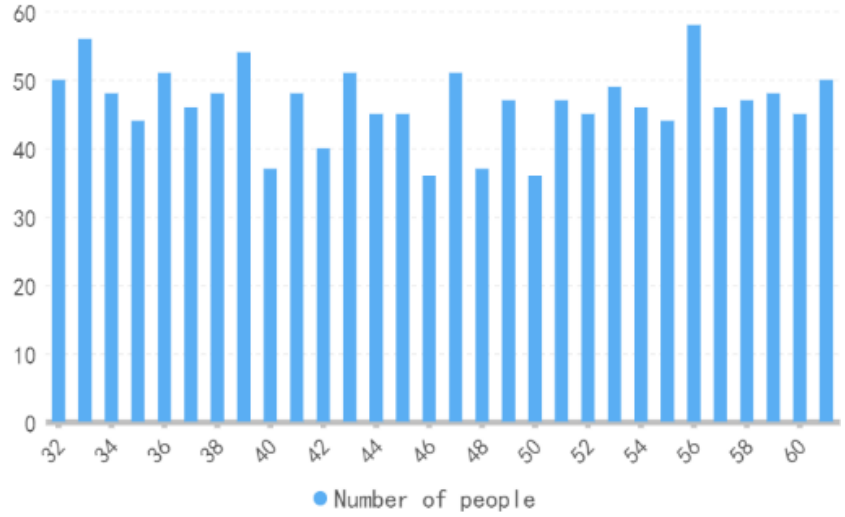

Figure 1. Histogram of patient age distribution

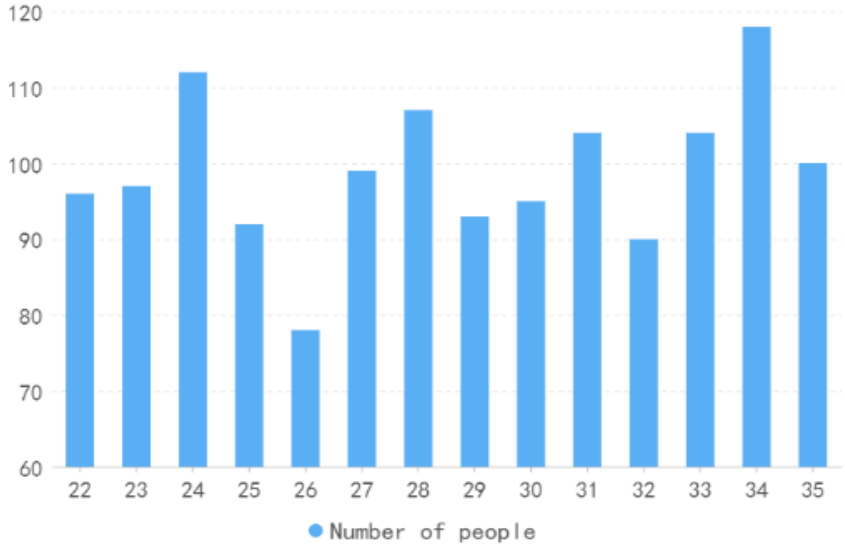

Figure 2. Histogram of patient BMI distribution

After the data was processed by means of deleting duplicate and error items and filling in missing values, 1174 valid data and 27 fields were retained. The feature descriptions are demonstrated in Table 1.

The missing rate of field RNA EF data in the data reached 27.5%. In order to avoid further data loss, some data were filled according to the similarity between the data. Before and after filling, the data balance was maintained at a good level.

The following sets are highly similar data filtered out (some data are not shown). Data in set A is leveraged to recover the data in set B. A= {1018717, 966926, 202928, 95479, 57257, 15, 3}, B= {1022090, 921626, 5, 5, 5, 13, 3}.

Table 1. Feature Description

Feature & Description	Feature & Description
Specific age	AST 1 aspartate transaminase ratio
Specific gender	ALT 1 alanine-transaminase-ratio for week1
BMI=height/(heavy^2)	ALT 4 alanine-transaminase-ratio for weeks4
Fever	ALT 12 alanine-transaminase-ratio for week12
Vomiting symptoms	ALT 24 alanine-transaminase-ratio for week24
Headache	ALT 36 alanine-transaminase-ratio for week36
Diarrhea symptoms	ALT 48 alanine-transaminase-ratio for week48
Fatigue and bone pain	ALT after 24 weeks alanine-transaminase-ratio
Jaundice symptoms	Basic virus RNA index
Epigastric pain	RNA index in week 4
WBC: Leukocyte index	RNA index in week 4
RBC: Red blood cell index	RNA end of the treatment
HGB: Hemoglobin index	RNA Elongation Facto
Platelet index	Baseline histological Grading
	Baseline histological staging

2.2. Adaboost

AdaBoost is the abbreviation of adaptive boost [8,9]. Its adaptability is to increase the weight of the samples incorrectly classified in the previous round in each iteration, and apply the increased weight to the next classifier for classification again. At the same time, Adaboost will add a new weak classifier every time it classifies and trains the model. Under the premise of preset values, the error rate is reduced continuously to form the latest strong classifier.

Algorithm flow is as follows. For the binary classification problem, the output is {- 1,1}. The weak classifier of round K is $G_k(x)$.

Step 1: Calculate the classification error of the weak classifier. The weighted error rate on the training set is

$$e_k = P(G_k(x_i) \neq y_i) = \sum_{i=1}^{m} w_{ki} I(G_k(x_i) \neq y_i) \quad (1)$$

Where the symbol I represents the indicator function.

Step 2: Calculation of Weak learning weight coefficient

$$\alpha_k = \frac{1}{2} log \frac{1-e_k}{e_k} \quad (2)$$

Step 3: Weight update of the next round of samples

$$w_{k+1,i} = \frac{w_{ki}}{z_k} exp(-\alpha_k G_k(x_i) y_i) \quad (3)$$

where z_k is the normalization factor, which makes the sum of sample weights of each training data set equal to 1.

$$z_k = \sum_{i=1}^{m} w_{ki}\, exp(-\alpha_k G_k(x_i) y_i) \quad (4)$$

When the sample is in the wrong classification, $G_k(x)y(i) = -1$, then $w_{k+1,i} > w_{ki}$ and the weight of the misclassified sample is increased; When the sample is in the correct classification $G_k(x)y(i) = 1$ then $w_{k+1,i} > w_{ki}$ and the weight of the correctly classified sample decreases.

Step 4: Combined with the strategy, the final classifier is constructed as follows:

$$G(x) = sign(\sum_{m=1}^{K} \alpha_m G_m(x)) \quad (5)$$

For this study, the Adaboost model can use different classification aslgorithms as weak classifiers, and cascade these weak classifiers. At the same time, Adaboost performs well in terms of accuracy ang takes full account of the weights of each weak classifier compared with other algorithm. The balanced data of this dataset can effectively avoid the disadvantage that Adaboost will reduce the accuracy of unbalanced data.

2.3. Other Models

In the research process of this paper, decision tree, random forest and LGBM are also selected as machine learning algorithms.

The study used decision tree model to predict the occurrence of abdominal liver pain in patients. Decision tree is mainly used to deal with classification problems. It is a tree structure that describes the process of sample classification. The construction of the decision tree starts from the root node and selects a feature of the sample. According to the test results, the samples are allocated to their child nodes. The samples are tested and allocated in recursion until the leaf node can allocate the samples to the class of the leaf node. At this point, a decision tree is formed.

In the research process, the decision tree establishes the relationship between output variables and input variables by training the data set, and then processes the data to be classified or predicted. [10].

Random forest, as its name implies, contains multiple decision trees. But his output still depends on a single decision tree. In calculation, it randomly forms a forest of multiple decision trees without any setting. For the classification algorithm, the decision tree in the forest will judge and classify the newly entered samples. Finally, the random forest will predict the attributes of the sample according to the category that has been selected the most times.

LightGBM, short for Light Gradient Boosting Machine, is the process of completing GBDT algorithm. It can conduct multiple trainings at the same time. Although the training time of LGBM is short and the memory required for training is small, it also maintains high accuracy. Compared with other models, it can process a large amount of data quickly without occupying too much memory.

3. Result

From Figure 1 and Figure 2 it could be observed that the patients are older and have higher BMI. Figure 3 shows that the number of people with various physical symptoms is between 660-700, indicating that there may be a link between several symptoms. The relationship between liver pain and other symptoms will be studied in the next study.

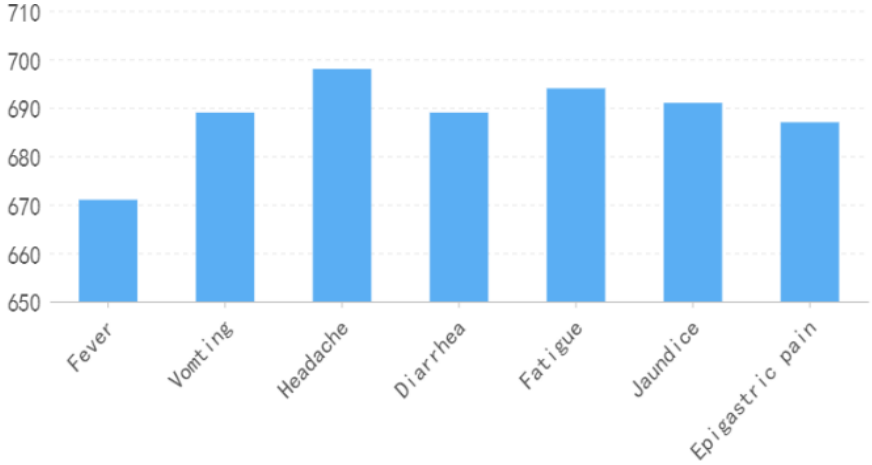

Figure 3. Statistical histogram of patient symptoms.

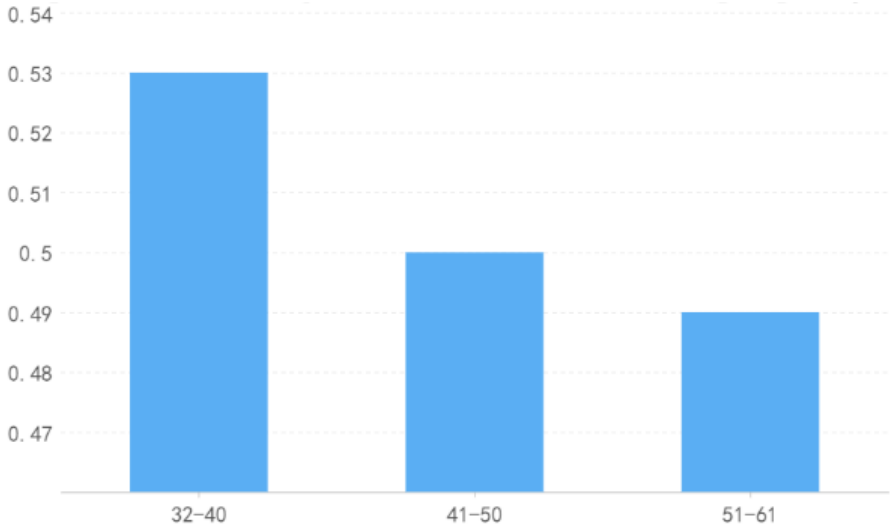

Figure 4. Obesity rate of different age groups

In the study, the age was divided according to the young, middle-aged and middle-aged. BMI greater than 28 is obesity. Figure 4 shows the jaundice rate and liver pain rate under different BMI, and the jaundice rate has a significant increase trend with the increase of BMI.

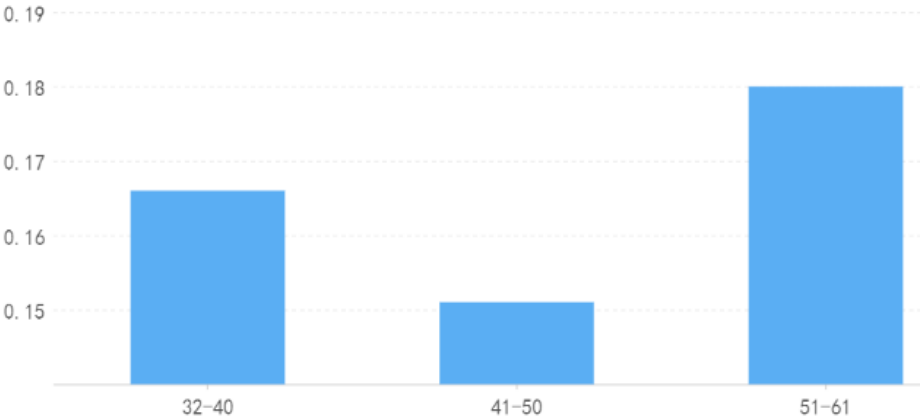

Figure 5. RNA at different ages mor than one million

Figure 5 shows the obesity rate and RNA greater than one million rates at different ages. RNA index is the viral RNA content of patients without treatment for 18 months. If the content is more than 1 million, it means that the virus replicates quickly in the patient. That is to say, the patient's physical condition is poor currently. RNA can indirectly reflect the replication rate of virus in human body. With the growth of age, the obesity rate decreases, and the indicators of physical condition fluctuate under the combined effect of the two.

The decision tree algorithm is used to predict whether patients will have liver pain, and the accuracy of the prediction result is about 58%. At the same time, carry out correlation analysis on the data. It proves that this field has poor correlation with other fields.

Table 2. Effectiveness Of Data Processing

	Adaboost	LGBM	Random forest	Decision tree
Before data processing	0.58	0.56	0.54	0.49
After data processing	0.72	0.66	0.65	0.62

Table 3. Classification Results

	Adaboost	LGBM	Random forest	Decision tree
0-precision	0.63	0.80	0.60	0.56
0-recall	0.89	0.62	0.64	0.66
0-f1-score	0.74	0.70	0.61	0.56
1-precision	0.73	0.59	0.70	0.70
1-recall	0.80	0.72	0.66	0.58
1-f1-score	0.76	0.65	0.68	0.63

The study used Adaboost, decision tree, LGBM and random forest to predict acute and chronic hepatitis C. The performances are demonstrated in Figure 2 and Figure 3. When the accuracy rate is not higher than 75%, priority shall be given to the accuracy rate. From the above two figures, Adaboost has the highest accuracy before and after data processing.

According to the data analysis, the higher the BMI and the older the age, the higher the probability of complications and the worse the physical condition of patients. However, relevant studies have proved that after a certain age, moderate overweight can improve their own resistance. Therefore, if the BMI of people over 45 years old is low, appropriate weight gain can be carried out to reduce the risk of disease and avoid excessive complications.

4. Conclusion

The main research content of this report is to classify the types of hepatitis C patients have, and analyze the influence of some factors in the treatment of hepatitis C. Based on the existing prediction of hepatitis C, this study proposes a classification prediction, which is an innovation in the auxiliary diagnosis of hepatitis C. In this study, four models in machine learning were compared when predicting the type of hepatitis C. Adaboost model performs better than other models in precision, recall and f1score, and achieves a high prediction accuracy. The study achieved 72% accuracy in predicting the type of hepatitis C through the Adaboost model. At the same time, the data used in the study, except the blood test data, can be perceived by the patients themselves, providing great convenience for the majority of patients when detecting the type of hepatitis C.

However, due to the lack of data in this study, many missing data cannot be completely filled, and the prediction accuracy is not excellent. In the future, based on the incremental prediction model, more data can be obtained in specific clinical trials to strengthen the existing model and update this study. Hepatitis C also has many complications, which can improve the accuracy of the model if more fields can be obtained.

References

[1] Aman, Wosen, et al. "Current status and future directions in the management of chronic hepatitis C." Virology Journal, vol. 9 (1), pp1 - 11, Dec. 2012.

[2] González-Grande, Rocío, et al. "New approaches in the treatment of hepatitis C." World journal of gastroenterology, vol. 22 (4), pp 1421, Jan. 2016.

[3] Hilgenfeldt EG, Schlacht Erman A, Firpi RJ. “Hepatitis C: treatment of difficult to treat patients.” World journal of hepatology, vol 7 (15), pp 1953, Jul. 2015.

[4] Sikavi C, Chen PH, Lee AD, Saab EG, Choi G, Saab S. “Hepatitis C and human immunodeficiency virus coinfection in the era of direct-acting antiviral agents: no longer a difficult to treat population.” Hepatology, vol 67 (3), pp 847 - 57, Mar. 2018.

[5] Kashif AA, Bakhtawar B, Akhtar A, Akhtar S, Aziz N, Javeid MS. “Treatment response prediction in hepatitis C patients using machine learning techniques.” International Journal of Technology, Innovation and Management (IJTIM), vol 1 (2), pp 79 - 89, Dec. 2021.

[6] Nasr M, El-Bahnasy K, Hamdy M, Kamal SM. "A novel model based on noninvasive methods for prediction of liver fibrosis." In2017 13th International Computer Engineering Conference (ICENCO), pp. 276 - 281, Dec. 2017.

[7] Anwar WA, El Gaafary M, Girgis SA, Rafik M, Hussein WM, Sos D, Mossad IM, Fontanet A, Temime L. "Hepatitis C virus infection and risk factors among patients and health-care workers of Ain Shams University hospitals, Cairo, Egypt." Plos one. vol 8; 16 (2), pp: e0246836, Feb. 2021.

[8] Ying C, Qi-Guang M, Jia-Chen L, Lin G. "Advance and prospects of AdaBoost algorithm." Acta Automatica Sinica, vol 1; 39 (6), pp 745 - 58, Jun. 2013.

[9] Chengsheng T, Huacheng L, Bing X. "AdaBoost typical Algorithm and its application research." InMATEC Web of Conferences, vol 139, pp 00222, 2017.

[10] Niuniu X, Yuxun L. "Notice of Retraction: Review of decision trees." In2010 3rd international conference on computer science and information technology, vol 5, pp 105 - 109, Jul 2010.