Adam Orłowski[1*]
[1]Computer Science Student at Faculty of Electrical Engineering,
Automatic Control and Informatics, Opole University of Technology
*Corresponding author: a.orlowski@student.po.edu.pl

Diagnosing Hepatitis C From Blood Tests Using Different Machine Learning Models

KEYWORDS

Hepatitis C; kNN; SVM; Logistic Regression; Random Forest

ABSTRACT

Hepatitis C is a chronic infectious disease that attacks the liver and is mainly transmitted through the bloodstream. Without proper treatment, it can lead to serious complications such as cirrhosis or even liver cancer. In the medical field, artificial intelligence (AI) is gaining importance all the time as a tool capable of increasing diagnostic efficiency and precision. With the passage of time, technological developments have increased the technical capabilities of computers to the point where processing large amounts of data and drawing conclusions based on them has become something widely available. This has contributed to a huge improvement in quality of AI solutions, dramatically increasing their impact on people's daily lives. In this article, we explore the potential of using machine learning to correctly diagnose an ongoing hepatitis C virus (HCV) infection. The goal of the paper is to develop and compare four binary classifiers: logistic regression, k-nearest neighbors (kNN), support vector machines (SVMs) and random forests, in order to select the one best suited to work with medical datasets. The learning process was conducted using blood test results from healthy donors and infected individuals. Metrics of accuracy and sensitivity were used to assess performance, the latter of which is particularly important so as not to miss any of the positive cases. In the end, among those that were tested random forest proved to be the best suited to medical applications such as diagnosing hepatitis C.

1. INTRODUCTION

Hepatitis C is an infectious disease transmitted primarily by the droplet route, which can lead to chronic, serious liver disease. Direct contact with human blood, activities such as intravenous drug use, blood transfusions, and medical procedures with unsanitary instruments are the main routes of transmission of this virus. In most cases, HCV infection remains asymptomatic and does not adversely affect the health of the carrier. However, without appropriate treatment, about 80% of infections can progress to a chronic phase, increasing the risk of developing cirrhosis and liver cancer [1].

Diagnosis of hepatitis C usually begins with screening for HCV antibodies. However, a positive result is not conclusive evidence of active infection, requiring additional testing for HCV RNA in serum. Making the diagnosis more difficult is the delayed production of antibodies by the infected person's body, which can take up to 8 weeks [2]. That is also not enough, if the infected are to be found as soon as possible.

Although the exact number of infected people worldwide is not known due to undiagnosed cases, an estimated 71 million people worldwide were living with chronic hepatitis C in 2015[3]. Since the WHO's introduction of guidelines to combat hepatitis C in 2016, there has been a downward trend in the number of new cases of chronic infections compared to previous years [3]. The WHO's goal is to reduce the incidence of disease by 90% before 2030, but these efforts have encountered obstacles related to the COVID-19 pandemic, significantly complicating the achievement of this goal [4].

Biochemical biomarkers, including aminotransferases (ALT, AST), gamma-glutamyl transferase (GGT) and bilirubin, are key in assessing liver status in patients with hepatitis C. Although they are not exclusive to this condition, high levels of ALT and AST may suggest hepatocellular damage, while GGT and bilirubin may indicate cholestasis or other liver dysfunction. Their importance as biomarkers in the diagnosis and monitoring of hepatitis C is well documented and a standard in clinical medicine [5, 6].

Binary classifiers are machine learning algorithms that categorize data based on two outcome classes, such as healthy or sick. They are the cornerstone of decision support systems, where their ability to learn and generalize from patterns in data allows them to predict and recognize patterns with high precision. These classifiers are particularly useful in conditions where the outcome is clearly defined and can be expressed as one of two categories, making them indispensable in many technological and scientific applications [7].

In medicine, and particularly in hepatitis C diagnosis, binary classifiers are used to analyze and interpret complex medical data, such as blood test results. By learning models from historical data, they can help identify patients at high risk for hepatitis C, contributing to early intervention and better treatment planning. Using binary classifiers in this domain can significantly increase the accuracy of diagnosis, minimizing the risk of false negative results, which is crucial in diseases with serious health consequences [8].

2. METHODOLOGY

The experiment used publicly available medical data [9]. The analyzed dataset consists of 615 blood samples, which were collected from both healthy donors and HCV-infected patients in various stages of the disease. Each sample is described by 14 attributes, as shown in Table 1. The samples are from individuals in the 19–77 age range and are characterized by a clear male preponderance and an almost ninefold preponderance of healthy individuals. Most of the samples represent healthy donors.

Table 1. Attributes in the dataset

Column	Definition
#	Study identifier
Category	Health status of the subject
Sex	Gender of the subject
Age	Age of the subject
ALB	Albumin concentration
ALP	Alkaline phosphatase
ALT	Alanine transaminase
AST	Aspartate transaminase
BIL	Bilirubin
CHE	Acetylcholinesterase
CHOL	Cholesterol level
CREA	Creatinine
GGT	Gamma-glutamyl transferase
PROT	Proteins

To ensure data consistency and integrity, missing values in the dataset were filled in using arithmetic averages appropriate for each attribute. In an effort to simplify the models and focus on the main purpose of the study, the *category* column, originally of the categorical type, was converted to a binary variable. This change was made to separate healthy from infected samples, regardless of the severity of the disease, allowing the focus to be solely on detecting the presence of disease.

The kNN, SVM, logistic regression and random forest algorithms were chosen for comparison. Such decisions were made because of their proven effectiveness in binary classification tasks, where it is necessary to balance generalization ability with robustness to overtraining. These methods are well suited to the limited size and wide variety of medical data, as highlighted in the [10].

The learning process used raw data, skipping standardization or normalization steps. Such a decision is based on the goal of evaluating the performance of the models in their most basic form. The entire experiment was conducted in Python and the default parameters available in the *scikit-learn* library were used for each model, allowing an unbiased comparison of their predictive abilities without additional optimization.

3. RESULTS

The calculations performed for the same input data yielded a range of results, summarized in Table 2. Comparison of the data in the table shows that all computational models achieved good accuracy. This does not mean that all of these results can be considered acceptable. The main reason for this conclusion is that if a true disease is misdiagnosed, the consequences can be fatal for the patient.

Table 2. Comparison of computational results

Computational Model	Accuracy, %	Sensitivity, %
Logistic Regression	89,43	54,17
KNN	88,62	50,00
Random Forest	93,50	66,67
SVM	87,80	45,83

The results discussed are graphically presented in Figure 1. Even a cursory analysis of the graphs indicates that the lowest diagnostic accuracy was observed for the SVM (support vector machine) model. In this case, the calculation accuracy is approximately 87.8%, and the sensitivity is 45.83%. The Random Forest model exhibited the best fit. For this model, the calculation accuracy is approximately 93.50%, and the sensitivity is 66.67%. For the other two models, Logistic Regression and KNN (k-nearest neighbors), the calculation results are similar, with calculation accuracy of approximately 89% and sensitivity of approximately 45.83%.

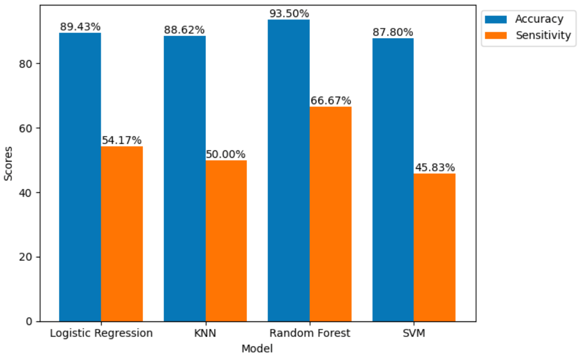

Fig. 1. Comparison of accuracy and sensitivity of all models

The computational methods discussed in this article and the results obtained from their application represent the first attempts to create efficient and effective tools to support the disease diagnosis process. Scientists still have many calculations and attempts to apply other computational algorithms ahead of them, but it is likely that such dynamic development of artificial intelligence will soon enable the creation of an efficient expert system to support the disease diagnosis process. This, in turn, will transform into more accurate disease diagnosis and the use of appropriate treatment methods, thus saving time and money.

4. CONCLUSIONS

The results show a clear advantage for random forest in hepatitis C diagnosis, both in terms of accuracy (93.5%) and sensitivity (66.67%), highlighting its utility in medicine. Logistic regression and KNN models, despite similar accuracy, show lower sensitivity, which may limit their effectiveness in identifying all cases of infection. The lowest SVM performance suggests that not all popular classification algorithms are equally effective in a specific medical context, highlighting the importance of choosing the right model. At the same time, it can be noted that the absolute sensitivity achieved by the tested methods is not satisfactory in a context as crucial as human health.

The current era is witnessing a rapid development of computational algorithms, simulation tools, and artificial intelligence. Therefore, further efforts will be directed towards finding increasingly accurate methods for diagnosing various diseases, including hepatitis C.

REFERENCES

[1] Alter M.J.: *Epidemiology of hepatitis C*, Hepatology, vol. 26, No. S3, 1997, pp. 62S–65S, DOI: 10.1002/hep.510260711,

[2] Chevaliez S., and Pawlotsky J.: *How to use virological tools for optimal management of chronic hepatitis C*, Liver International, vol. 29, No. s1, 2009, pp. 9–14, DOI: 10.1111/j.1478-3231.2008.01926.x,

[3] World Health Organization: *Global health sector strategies on HIV, viral hepatitis and STIs for 2016–2021*, 2016,

[4] Blach S., et al.: *Impact of COVID-19 on global HCV elimination efforts*, Journal of Hepatology, vol. 74, No. 1, 2021, pp. 31–36, DOI: 10.1016/j.jhep.2020.07.042,

[5] Dooley J.S., Lok A.S.F., Burroughs A.K., and E.J. Heathcote E.J.: *Eds., Sherlock's Diseases of the Liver and Biliary System*. Wiley, 2011. doi: 10.1002/9781444341294,

[6] Marcellin P.: *Hepatitis C: the clinical spectrum of the disease*, Journal of Hepatology, vol. 31, 1999, pp. 9–16, DOI: 10.1016/S0168-8278(99)80368-7,

[7] James G., Witten D., Hastie T., and Tibshirani R.: An Introduction to Statistical Learning, vol. 103. *New York, NY: Springer New York,* 2013, DOI: 10.1007/978-1-4614-7138-7,

[8] Kuhn M., and Johnson K.: Applied Predictive Modeling. *New York, NY: Springer New York*, 2013, DOI: 10.1007/978-1-4614-6849-3,

[9] Lichtinghagen R., Klawonn F., and Hoffmann G.: 2020, *HCV Data*, University of California, Irvine, DOI: 10.24432/C5D612,

[10] Couronné R., Probst P., and Boulesteix A.-L.: *Random forest versus logistic regression: a large-scale benchmark experiment*, BMC Bioinformatics, vol. 19, No. 1, 2018, p. 270, DOI: 10.1186/s12859-018-2264-5.