Artificial Intelligence in Industry 4.0: The future that comes true: AI

Karabegović, Isak; editor

2024-09-17

https://bastina.anubih.ba/handle/123456789/791

DOI: 10.5644/PI2024.215.16

Liver Disease Classification Using Machine Learning

Madžida Hundur Hiyari[*1], Nejra Merdović[1], Faruk Bećirović[1], Emina Mrđanović[1], Adna Softić[1]

Abstract: *Hepatitis C virus (HCV) is a significant cause of liver-related diseases including acute and chronic hepatitis, cirrhosis, and hepatocellular carcinoma. Despite the availability of advanced treatments, underdiagnosis remains a critical challenge, particularly in resource-limited settings. This study explores the application of machine learning algorithms, specifically the K-Nearest Neighbors (KNN) method, to enhance the diagnosis of HCV by classifying patients into healthy, potentially diseased, and diseased categories based on liver function test results. Using a biomedical dataset of 615 patients, the model achieved high accuracy (99%), precision (98%), and sensitivity (99%), indicating its potential effectiveness in identifying HCV-infected individuals. The study highlights the importance of feature selection in improving model performance and discusses the implications of the findings for enhancing HCV diagnosis and management.*

Keywords: *Hepatitis C virus (HCV), machine learning, K-Nearest Neighbors (KNN), classification*

1. Introduction

There are many microbes, toxins, autoimmune diseases and neoplastic diseases that cause liver inflammation, but 5 viruses (hepatitis A, B, C, D, and E) cause liver disease as their main pathogenesis [1]. The hepatitis C virus (HCV) as a positive-sense, single stranded RNA flavivirus with seven known subtypes represents the main cause of acute and chronic hepatitis, cirrhosis, and hepatocellular carcinoma [2]. It is estimated that 71 million people worldwide are infected with the virus, which is leading us to an estimated prevalence of 0.64% of the total population of the European Union. This virus is usually transmitted by blood or body fluids, through sharing needles, blood transfusion or by transmission to infants born to HCV viraemic mothers [3]. HCV diagnosis must be as simple as possible and should be linked to the HCV process. There are different categories of laboratory tests: classical serologic tests able to detect anti HCV antibodies (Ab) cells indirect tests, and assays that can detect and

[*1]Verlab Research Institute for Biomedical Engineering, Medical Devices and Artificial Intelligence, Sarajevo, Bosnia and Herzegovina
E-mail: madzida@verlabinstitute.com; adna@verlabinstitute.com

quantify HCV particle components such as HCV RNA or HCV Core Antigen (cAg) named direct tests [4]. These tests are complementary for the diagnosis of infection, therapeutic decisions and to assess treatment properly. Unfortunately, the underdiagnosis of HCV remains a major obstacle to attaining global eradication, despite advances in therapy [5]. A major challenge is the low diagnosis rate, especially in nations with little resources, which is mostly caused by the expensive price of necessary molecular diagnostic instruments. Furthermore, because extensive HCV testing is less popular than it is for HIV, people infected with chronic hepatitis C (CHC) frequently remain unaware of their infection due to the mild clinical signs associated with the illness [6].

Machine learning (ML) algorithms are adept at analysing medical phenomena by capturing complex and nonlinear relationships in clinical data. The ML algorithms, such as classification techniques, can be utilised to develop a model to diagnose HCV by identifying people who are infected with the virus. However, inappropriate characteristics in the attribute set can spoil the classifier's performance [7]. Feature selection defines a subset of features or variables that describe data to obtain a more compact and essential representation of the available information and ignore all other redundant and irrelevant features [8]. Feature selection is a powerful way to enhance the functioning and reduce the model development time of a classifier.

2. Methods

In the context of supervised machine learning, algorithms are trained on labeled datasets, with each training example consisting of an input and a correspondingly correct output. The main goal is to deduce a function that connects inputs to intended outputs so that the model can forecast new, unobserved data with accuracy. Choosing a suitable model, training it on the labeled data, and assessing its performance with a different test set are the steps in this learning process. The two primary tasks of supervised learning are regression, which anticipates continuous values, and classification, which predicts categorical labels. Its applications cover a wide range of fields, including natural language processing, picture identification, and medical diagnostics, demonstrating its adaptability and crucial significance in contemporary data-driven decision-making. [9]

For classification and regression applications, the K-Nearest Neighbors (KNN) method is a straightforward yet effective supervised learning method. It functions based on the idea that comparable data points should produce comparable results. When a new input is given, KNN finds the 'k' nearest data points in the training set, usually with the use of a distance metric like Euclidean

distance, then predicts the output based on these neighbors' average value (for regression) or majority class (for classification). A critical hyperparameter influencing algorithm performance is 'k'; a small value of 'k' may cause overfitting, while a big value of 'k' may cause the model to be oversimplified [10].

KNN stands out for being straightforward, simple to use, and effective, especially in situations where the decision boundary is wildly erratic. However, because it needs to calculate the distances to every training sample for every prediction, it can be computationally demanding, especially with large datasets. In spite of this, KNN is still often utilised because of its ease of use and capacity to manage multiclass classification issues. In the current era of big data, KNN methods offer an especially effective approach for identifying valuable patterns and creating case-based reasoning algorithms for artificial intelligence (AI). [11]. Based on the structure of the dataset and the data within it KNN is chosen for the classification task due to its effectiveness in similar tasks.

The dataset used in this study is from the field of biomedicine and tracks the health status of 615 patients who underwent blood tests specifically to check liver function. Within this group, 533 patients had good blood profiles and normal liver function, categorized as 0, while the remaining 82 patients were sent for additional tests due to suspected liver disease, categorized as 1, or immediately confirmed with liver disease, categorised as 2. The dataset consists of: identification number (ID), category label (Category), age (Age), gender (Sex), and measurements of 10 different blood components. To adequately prepare the data for machine learning applications, all missing values (total of 31) were replaced with the mean value of the respective column. The data is split into two groups: training data (80% of the dataset) and testing data (20% of the dataset).

The goal of the study is to classify patients into healthy, potentially diseased, and diseased categories. The chosen classifier is the K-nearest neighbors (KNN) due to its simplicity and quick execution. Table 1 lists the selected parameters of this classifier. N-neighbors is a crucial parameter for this method, defining the number of nearest neighbors considered when making decisions. This parameter needs careful selection as it impacts the model's complexity and susceptibility to overfitting or underfitting. In this study, a value of 5 neighbors was chosen, which was shown to be optimal during testing. The "weights" parameter refers to

the weighting function used in prediction. In this study, a uniform function was selected, meaning all neighbors have equal weight in decision-making. This implies that all neighbors are considered equally regardless of their distance from the query sample. The distance measure chosen is the commonly used Euclidean distance. Using Equation 1, it measures the straight line between the input point and the point being measured.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} \quad (y_i - x_i)^2} \qquad (1)$$

Table 1. Selected parameters for KNN classifier

Method	**K – nearest neighbors**
n_neighbors	5
weights	uniform
metric	Euclidean

3. Results and Discussion

After training and testing the created machine learning model, evaluation of the results was conducted using the parameters: accuracy, precision, sensitivity (Table 2), and analysis of the confusion matrix (Figure 1).

Table 2. Value of parameters used for evaluation of the created machine learning model

Parameters	**Percentage**
Accuracy	99%
Precision	98%
Recall	99%

An accuracy of 0.99 was achieved, indicating a high proportion of correctly classified instances among the total number of instances. However, due to the

imbalanced classes in the dataset used, drawing conclusions based solely on accuracy is not sufficient to assess the model's performance. Therefore, further analysis was carried out.

A precision of 0.98 was achieved, indicating a low rate of false positive classifications, which is also evident when observing the confusion matrix from Figure 1. Sensitivity is another significant parameter for analyzing model performance, and in this study, a sensitivity of 0.99 was achieved. This indicates that the model identifies nearly all actual positive instances.

These results indicate that the combination of the uniform weighting function and Euclidean distance measure provides an effective and reliable means of decision-making.

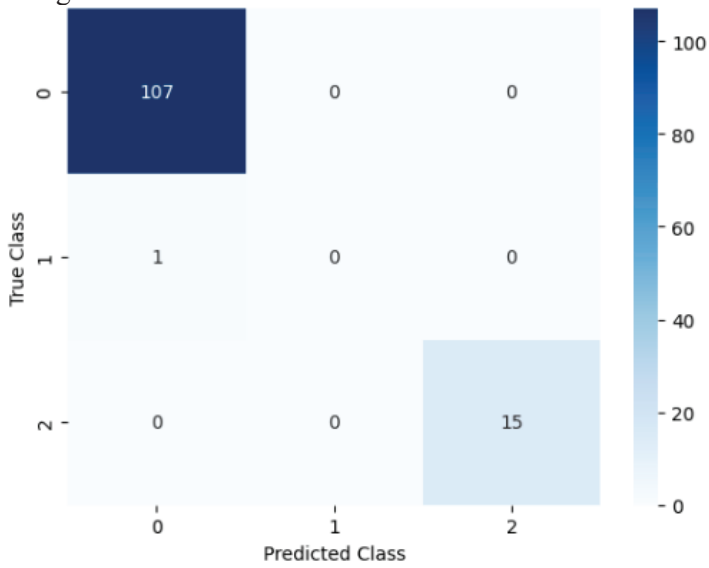

Figure 1. Confusion Matrix

From the confusion matrix, we can see that:

- 107 instances were correctly classified as class 0 (healthy patients),
- 1 instance that belongs to class 1 (potentially diseased patients) was incorrectly classified as class 0 (healthy patients),
- 15 instances were correctly classified as class 2 (diseased patients).

There were very few instances classified as false positives or false negatives, with the majority of instances being correctly classified. Based on all the above, it can be concluded that the model's performance is satisfactory.

To further understand the model's robustness, additional metrics, including F1-score and the Area Under the Receiver Operating Characteristic (ROC-AUC) curve were calculated. The F1-score, which considers both precision and recall, was 0.99, demonstrating balanced performance across both metrics. ROC-AUC scores for the model are as follows: 0.97 for class 0, 0.5 for class 1, and 1.0 for class 2, with an average of 0.99 (Figure 2). These results suggest that the model performs exceptionally well for classes 0 and 2, but has limited discriminatory ability for class 1, indicating potential areas for improvement in classifying potentially diseased patients.

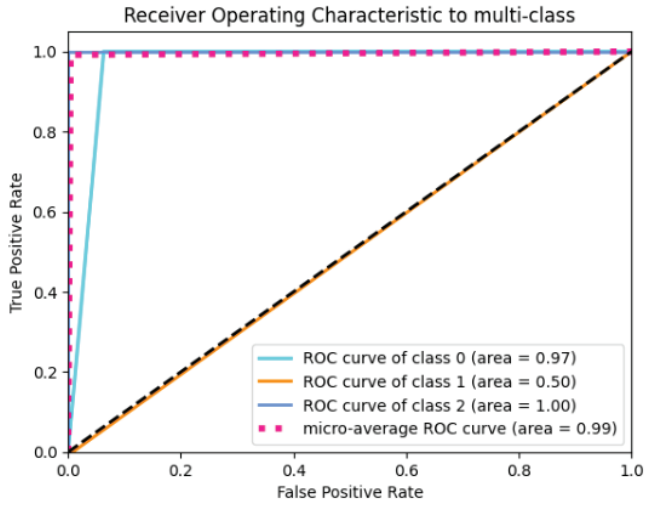

Figure 2. Receiver Operating Characteristic to multi-class

At the end, overfitting of the machine learning model was checked using cross-validation. It was ensured that each subset (10-fold in this case) of cross-validation was representative of the entire dataset in terms of class distribution. An average accuracy of 0.99 was obtained through cross-validation. Since this average accuracy is equal to the accuracy achieved on the test dataset, it can be concluded that the model did not overfit. In other words, the model generalizes well to new data.

4. Conclusion

This study confirms the effectiveness of the K-Nearest Neighbors (KNN) algorithm in classifying patients based on liver function test results. Using a dataset of 615 patients, we achieved high performance metrics: 99% accuracy, 98% precision, and 99% sensitivity. The low rates of false positives and false negatives, validated by the confusion matrix, further support the model's reliability.

Despite class imbalance, the KNN model demonstrated robustness and generalizability, as verified by cross-validation. These results highlight KNN's suitability for medical diagnostics, particularly in multiclass classification tasks. Future research could enhance KNN's performance by integrating advanced techniques, expanding its applicability in biomedicine and beyond. Specifically, investigating alternative weighting functions to assign greater importance to closer neighbors could potentially improve classification accuracy. Additionally, evaluating different distance metrics or hybrid approaches may capture more relevant similarities in high-dimensional or non-linear data spaces. Expanding the dataset to include larger and more diverse samples will help validate the algorithm's applicability across various liver diseases.

Integrating KNN with other machine learning techniques and ensemble methods could further enhance diagnostic performance and robustness, making the approach more effective in clinical settings.

5. References

[1] Prasidthrathsint K, Stapleton JT. Laboratory Diagnosis and Monitoring of Viral Hepatitis. Gastroenterol Clin North Am. 2019 Jun;48(2):259-279. doi: 10.1016/j.gtc.2019.02.007. Epub 2019 Apr 1. PMID: 31046974; PMCID: PMC10461253.

[2] Duncan, J.D.; Urbanowicz, R.A.; Tarr, A.W.; Ball, J.K. Hepatitis C Virus Vaccine: Challenges and Prospects. Vaccines 2020, 8, 90. https://doi.org/10.3390/vaccines8010090

[3] Abu-Freha N, Mathew Jacob B, Elhoashla A, Afawi Z, Abu-Hammad T, Elsana F, Paz S, Etzion O. Chronic hepatitis C: Diagnosis and treatment made easy. Eur J Gen Pract. 2022 Dec;28(1):102-108. doi: 10.1080/13814788.2022.2056161. PMID: 35579223; PMCID: PMC9116263.

[4] Shahid I, Alzahrani AR, Al-Ghamdi SS, Alanazi IM, Rehman S, Hassan S. Hepatitis C Diagnosis: Simplified Solutions, Predictive Barriers, and Future

Promises. Diagnostics. 2021; 11(7):1253. https://doi.org/10.3390/diagnostics11071253

[5] Sonia Arca-Lafuente, Paula Martínez-Román, Irene Mate-Cano, Ricardo Madrid, Verónica Briz, Nanotechnology: A reality for diagnosis of HCV infectious disease, Journal of Infection, Volume 80, Issue 1, 2020, Pages 8-15, ISSN 0163-4453, doi: 10.1016/j.jinf.2019.09.010.

[6] Roger S, Ducancelle A, Le Guillou-Guillemette H, Gaudy C, Lunel F. HCV virology and diagnosis. Clin Res Hepatol Gastroenterol. 2021 May;45(3):101626. doi: 10.1016/j.clinre.2021.101626. Epub 2021 Feb 23. PMID: 33636428.

[7] John GH, Kohavi R, Pfleger K (1994) Irrelevant features and the subset selection problem. Machine learning proceedings. Elsevier, Amsterdam, pp 121–129

[8] Triantaphyllou E, Felici G (2006) Data mining and knowledge discovery approaches based on rule induction techniques. Springer, New York.

[9] Shruthi H. Shetty, Sumiksha Shetty, Chandra Singh, Ashwath Rao. Supervised Machine Learning: Algorithms and Applications. February 2022. doi: 10.1002/9781119821908.ch1.

[10] K. Taunk, S. De, S. Verma and A. Swetapadma. A Brief Review of Nearest Neighbor Algorithm for Learning and Classification. 2019 International Conference on Intelligent Computing and Control Systems (ICCS), Madurai, India, 2019, pp. 1255-1260, doi: 10.1109/ICCS45141.2019.9065747.

[11] S. Zhang, "Challenges in KNN Classification," in IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 10, pp. 4663-4675, 1 Oct. 2022, doi: 10.1109/TKDE.2021.3049250.