

Seeking best performance: a comparative evaluation of machine learning models in the prediction of hepatitis C

Michael Cabanillas-Carbonell¹, Joselyn Zapata-Paulini²

¹Faculty of Engineering, Universidad Privada del Norte, Lima, Peru

²Graduate School, Universidad Continental, Lima, Peru

Article Info

Article history:

Received Mar 5, 2024

Revised Jan 22, 2025

Accepted Mar 25, 2025

Keywords:

Evaluation

Hepatitis

Machine learning

Models

Prediction

ABSTRACT

Hepatitis C is a disease that affects millions of people worldwide. It is spread through contact with contaminated blood through injections, transfusions, or other means. It is estimated that with early detection patients have a higher rate of recovery. The objective of this study is to perform a comparative evaluation of different models focused on the prediction of hepatitis C, to determine which of the models offers better performance in accuracy, precision, and sensitivity. The models used were logistic regression (LR), random forest (RF), K-nearest neighbors (KNN), decision tree (DT), and gradient boosting (GB), aimed at hepatitis C prediction. The training of the models was carried out using a dataset composed of 615 records, which incorporate 14 attributes. The structure of the article is divided into six sections, including introduction, review of related articles, methodology, results, discussion, and conclusions. The performance of the models was evaluated through metrics such as accuracy, sensitivity, F1 count, and, mainly, precision. The results obtained place the DT model as the most efficient predictor, reaching a precision, accuracy, sensitivity, and F1-score of 95%.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Joselyn Zapata-Paulini

Graduate School, Universidad Continental

Alfredo Mendiola 5210, Los Olivos 15311, Lima, Perú

Email: 70994337@continental.edu.pe

1. INTRODUCTION

Hepatitis C (HCV) is a viral disease that was historically classified simply as viral hepatitis when it was not identified as type A or type B. This virus is mainly transmitted through blood transfusions and other contacts with contaminated blood. Once the infection is acquired, patients face a higher risk of developing chronic liver diseases, such as hepatocellular carcinoma or cirrhosis [1]. According to the World Health Organization (WHO), approximately 58 million people worldwide are chronically infected with HCV, and more than 1.2 million new infections are reported each year. Of these, about 3 million children and adolescents are also chronically affected [2]. HCV primarily attacks liver cells and is unique to humans. This virus possesses a remarkable ability to evade both innate and adaptive immunity, resulting in chronic infections in approximately 70% of cases [3]. The appearance of HCV antibodies is a common indicator of infection; according to records, Africa and Asia are known to be the continents with the highest prevalence rates of these antibodies, while Australia, North America, and Western Europe show the lowest rates [4]. Although preventive methods, such as vaccination and the use of promising new drugs, can cure HCV infection in up to 70% of treated patients [5], most infected individuals are unaware of their condition, so it is a priority to implement screening programs for early and timely detection of the disease [6].

HCV is divided into a total of seven genotypes, which are divided into multiple subtypes, the classification of these genotypes depends on the ethnic group and mode of transmission [7]. Genotype 1 is the most common with 46% of all cases, with presence in regions of Asia, North America, Australia, South America, Northern and Western Europe Northern and Western Europe [8], [9]. Genotypes 2, 4, and 6 account for most of the remaining HCV cases, but only one case of type 7 has been reported so far in Canada [10]. During a prevalence review in the United States, an average of 3.5 million people in the country were identified as being infected with HCV [11]. It was also identified that 57% of people had been screened and were aware of their condition, and 50% had HCV antibodies in their system [12]. The prevalence of the disease in countries such as Egypt is 18% to 22%, in Italy it is 2.5% to 10%, in Pakistan it is 4.9%, in China it is 3.2% and in Indonesia, it is 2.1% [13].

Currently, artificial intelligence (AI) methods, such as machine learning (ML) and deep learning (DL) models, are playing a crucial role in the process of diagnosis, prediction, and treatment of diseases, such as diabetes, Alzheimer's disease, and heart disease [14], [15]. In ML-related studies, algorithms or models are employed to identify patterns or indicators within large data sets [16], [17]; to detect the possible existence or absence of the ailment under investigation [18]. Therefore, this tool can be useful for the development of an HCV prediction model.

This study aims to address the need to develop innovative techniques to predict HCV infection. Through the benchmarking of various ML models, in order to determine which of the models offers better performance in accuracy, precision, and sensitivity. To this end, the logistic regression (LR), random forest (RF), K-nearest neighbors (KNN), decision tree (DT), and gradient boosting (GB) models are conceptualized and developed. This research aims not only to facilitate the early detection of HCV, but also to improve the design of more effective treatments against the virus, thus contributing to the reduction of the overall impact of this disease.

This article is structured in six parts. The first part details and contextualizes the problems of the study. The second part is a review of related studies. In the third part, we develop the methodology divided into two sections, in the first section we conceptualize the ML models, and in the second section, we develop the case study. In part four of the article we present the results of the models. In part five we discuss the results obtained with related studies. Finally, in part six we present the conclusions.

2. RELATED WORK

In this section, we discuss work related to the case study. Alizargar *et al.* [19], aimed to use different ML models to predict hepatitis C with blood tests, to treat patients in the early stages of infection; in their methodology, they used data mining techniques to process the datasets, to subsequently train six ML models; the study concluded that the support vector machine (SVM) and extreme gradient boosting (XGBoost) models reached an accuracy of 0.82, being the best results achieved. Likewise, in the study from Syafaah *et al.* [20] they evaluated the level of accuracy achieved by different ML models to determine which is the most accurate in the detection of hepatitis C; in their methodology they took into account multiple indicators of blood tests to detect the disease, to subsequently train the classification models; the results of the study positioned neural networks (NN) as the best with 0.9512 in accuracy, followed by KNN, Naive Bayes (NB) and RF with 0.8943, 0.9024, and 0.9431, respectively. On the other hand, in the study from Ma *et al.* [21] they evaluated several ML classifiers for early prediction of hepatitis C; in their methodology, they used the blood records of multiple patients diagnosed with this disease to train the models; the study positioned the XGBoost model as the best in predicting the disease with an accuracy of 0.9156, precision of 0.98 and sensitivity of 0.98. In turn, Ahammed *et al.* [22] they sought to classify the liver states of people infected with the virus by making use of three ML models; in their methodology, they employed the dataset from the ICU repository, which was subjected to the synthetic minority oversampling technique (SMOTE), and subsequently applied feature selection methods to finally train the models; the study concluded that the KNN model achieved the best performance with 0.9440 in accuracy. In a real case, Farghaly *et al.* [23] evaluated different ML models focused on predicting hepatitis C, in healthcare workers in Egypt; for training the models they employed a two-stage dataset, in the first stage the dataset was without feature selection and in the second stage they applied feature selection focused on identifying forward sequences; the study concluded that the RF model achieved the best result since in the first stage it reached an accuracy of 0.9406 and in the second stage a 0.9488. In the study from Ali *et al.* [24] they analyzed and evaluated the performance of multiple ML algorithms for the early diagnosis of hepatitis C; in their methodology, they applied processing techniques on the dataset such as feature selection, forward feature selection and SMOTE; the results of the study specified that the evaluated models achieved an average accuracy of 0.83, such as KNN, RF, and LR models with a performance of 0.831, 0.824 and 0.829, respectively. On the other hand, Chen *et al.* [25] they propose a unique model for each of the patients seeking to be diagnosed with hepatitis C; the results position the XGBoost model as the best with 0.95 in accuracy and 0.70 in sensitivity. Santos [26]

Seeking best performance: a comparative evaluation of machine ... (Michael Cabanillas-Carbonell)

contrast different ML models for the prediction of the severity of hepatitis C infection in patients; in their methodology, they used different data preprocessing techniques, data engineering, and hyperparameter optimization applied to both the dataset and the four algorithms that were evaluated; the study concluded that the RF and GB models achieved the best accuracy and precision with 0.9350. Similarly, Harabor *et al.* [27], they developed a study to compare and evaluate the performance of four ML models for the prediction of Hepatitis B and C status; the results of the study showed that the model with the best predictive performance is KNN, with an accuracy of 0.981, followed by SVM and RF with equal accuracy of 0.976 and NB with 0.957. The study from El-Salam *et al.* [28] aimed to analyze and evaluate different ML models for early prediction of hepatitis C; in their methodology, they applied different techniques such as feature selection for data processing; the results of the study positioned the Bayesian Network model with the best performance with 0.748 in accuracy. Hashem *et al.* [29] contrasted different ML models focused on the prediction of liver fibrosis in patients with chronic hepatitis C; the results determined that the models obtained results ranging from 0.663 to 0.844 in accuracy. Kareem [30] four ML models to classify and diagnose hepatitis C; the results of the study positioned DT with the best performance with an accuracy of 0.9344. Meanwhile, Lilhore *et al.* [31] they propose a hybrid model between RF and SVM for the prediction and classification of hepatitis C; in their methodology, they employed various optimization techniques for the models and SMOTE to create synthetic data to enhance the dataset; the study concluded that the hybrid model achieved an accuracy of 0.9589. Finally, Ghazal *et al.* [32] used the SVM model for hepatitis C prediction; the study concluded that the model managed to achieve an accuracy of 0.979.

3. METHOD

In this section of the study, we present the methodology divided into two parts, in part A, we conceptualize the ML models (LR, RF, KNN, DT, and GB) that we employ in this study. In part B, we develop the case study by analyzing and optimizing the dataset to subsequently train the models.

3.1. Description of the ML models

3.1.1. Logistic regression

LR is used in ML for binary classification, for example, to predict the presence or absence of a disease in a patient, this is done by analyzing a dataset that includes several features and a target variable [33]. The model is a supervised learning algorithm, which aims to model the relationship between input features and output labels, consequently, the result is expressed as the probability that the input belongs to a particular class [34]. Unlike other models, LR has some limitations such as assuming that the input features and output labels are linear and the features independent, to overcome these drawbacks other models were created [35]. In (1) the model is mathematically represented. Y is the variable representing the probability of an event occurring, denoted by $P(Y)$.

$$P(Y) = \frac{1}{1 + e^{-(b_0 + b_1X_1 + b_2X_2 + \dots + b_nX_n)}} \quad (1)$$

3.1.2. Random forest

RF represents a general ML algorithm that is used in both classification and regression tasks [36]. This ensemble learning method creates multiple decision trees during training and generates a modal class (in classification) or an average prediction (in regression) from the individual trees [37]. Each tree in the forest is created from a random selection of training data and features, this introduction of randomness helps to reduce overfitting and improve the accuracy of the model [38]. In (2) shows the formula that the model uses to estimate the predictions for each tree.

$$\bar{r}_n(X, D_n) = E_\theta[r_n(X, \theta, D_n)] \quad (2)$$

3.1.3. K-nearest neighbors

KNN in ML is used in both classification and regression, it is based on clustering data points into groups and assigning them to the group containing the closest data point, called k-nearest neighbor [39]. Moreover, it makes no assumptions about the distribution of the data, as it is a nonparametric model [40]. The model uses the Euclidean equation, represented in (3), to calculate the distance between continuous variables, while it resorts to the overlap metric for discrete variables when measuring the proximity between neighbors [41].

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^p (x_{ri} - x_{rj})^2} \quad (3)$$

3.1.4. Decision tree

The DT model is presented as a supervised ML algorithm used in classification and regression tasks [42]. It works by recursively dividing the data into subsets according to the most relevant attribute, creating a tree structure, this process continues until the data in each subset becomes consistent concerning the target variable or until a predetermined stopping criterion is met [43]. In (4) the mathematical equation of the model is expressed. Within the equation, P_n is used to express the probability of non-occurrence, s is used to represent the sample, E is interpreted as the entropy and P_y is used to express the probability of occurrence.

$$E(s) = \sum_{k=0}^n \binom{n}{k} - P_y * \log 2P_n \quad (4)$$

3.1.5. Gradient boosting

Is an ensemble learning technique that integrates the predictions of various base estimators, usually DT-based, to increase model accuracy and robustness [44]. In modeling, the word ‘gradient’ refers to the implementation of a gradient descent algorithm to minimize losses when integrating new models into an ensemble [45]. Likewise, the term ‘boosting’ is used to describe the progressive inclusion of models in an ensemble, with the particularity that each new model has the function of correcting the errors of its predecessors [46]. As a result, a powerful predictive model is obtained that can identify complex patterns in the data and has a lower tendency to overfit [47]. The model equation can be expressed in (5). Where $f(x)$ represents the prediction function, $h(x)$ corresponds to the prediction of the i -th least robust model, \hat{y} denotes the final model accuracy, and γ is the learning coefficient.

$$\hat{y} = f(x) = \sum \gamma * h(x) \quad (5)$$

3.2. Case study

3.2.1. Understanding the dataset

A dataset extracted from the UCI ML repository was used for the ML model training process. This dataset contains laboratory values of blood donors, patients with hepatitis C, and demographic values. It has 615 records and 14 attributes, where all are numerical, except category and sex. The attributes are the following: “X” (patient id), “Category” which refers to the diagnosis (values: ‘0=Blood donor’, ‘0s=suspected blood donor’, ‘1=Hepatitis’, ‘2=Fibrosis’, ‘3=Cirrhosis’), “Age” (in years), “Sex” (h,m) and the laboratory attributes: “ALB” (albumin blood test), “ALP” (Alkaline Phosphatase), “ALT” (Alanine Transaminase), “AST” (Aspartate Transaminase), “BIL” (Bilirubin), “CHE” (Acetylcholinesterase), “CHOL” (Cholesterol), “CREA” (Creatinine), “GGT” (Gamma-Glutamyl Transferase), “PROT” (Protein). The development process of the study is detailed in Figure 1.

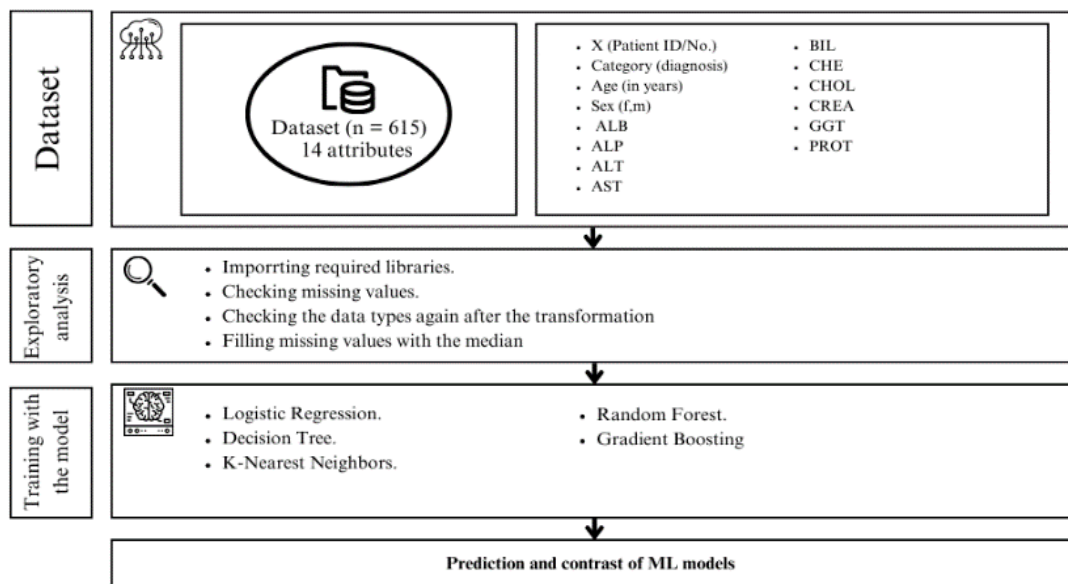


Figure 1. Case study development process

3.2.2. Data preparation

Before the exploratory analysis of the data, we performed a general analysis of the characteristics of the attributes contained in each variable. After loading the data set, we noticed the existence of a column called “Unnamed: 0”, which we proceeded to eliminate. After this, we verified the type of data stored in each column of the dataset, we noticed that the columns “category” and “age” are of type object, so we proceeded to transform them to type int to ensure a better processing of the data by the models, the results can be seen in Table 1. Likewise, we verified the unique values and the existence of missing values, identifying the columns “ALP”, “ALT”, and “PROT” with missing elements and proceeded to fill these values. Table 2 shows the final result of the data set.

Table 1. Data types

Attribute	Type
Category	int64
Age	int64
Sex	int64
ALB	float64
ALP	float64
ALT	float64
AST	float64
BIL	float64
CHE	float64
CHOL	float64
CREA	float64
GGT	float64
PROT	float64
dtype: object	

Table 2. Content the data set

	Category	Age	Sex	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT
0	0	32	0	38.5	52.5	7.7	22.1	7.5	6.93	3.23	106	12.1	69
1	0	32	0	38.5	70.3	18	24.7	3.9	11.17	4.8	74	15.6	76.5
2	0	32	0	46.9	74.7	36.2	52.6	6.1	8.84	5.2	86	33.2	79.3
3	0	32	0	43.2	52	30.6	22.6	18.9	7.33	4.74	80	33.8	75.7
...
611	1	64	1	24	102.8	2.9	44.4	20	1.54	3.02	63	35.9	71.3
612	1	64	1	29	87.3	3.5	99	48	1.66	3.63	66.7	64.2	82
613	1	46	1	33	68.28392	39	62	20	3.56	4.2	52	50	71
614	1	59	1	36	68.28392	100	80	12	9.07	5.3	67	34	68

3.2.3. Exploratory analysis of the data

In Figure 2, an exhaustive analysis of the target variable “Category” was carried out. The results reveal that approximately 70% of the patients show signs of hepatitis C, while the remaining 30% show a health status considered normal. This finding suggests a significant prevalence of hepatitis C in the studied population and a significant imbalance that will have to be taken into account when training ML models.

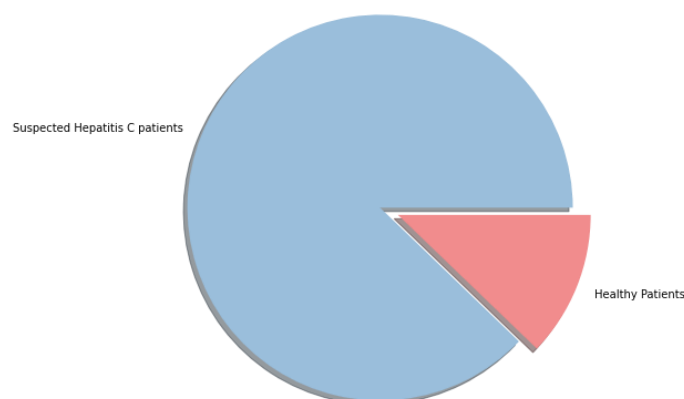


Figure 2. Target variable

On the other hand, when performing a univariate analysis of the patient data, a significant disparity in terms of gender was observed. In Figure 3, the results reveal that 61.30% of the individuals are male, while 38.70% correspond to the female sex in the data set. This finding highlights a marked predominance of males in the sample, which could have important implications when training the models.

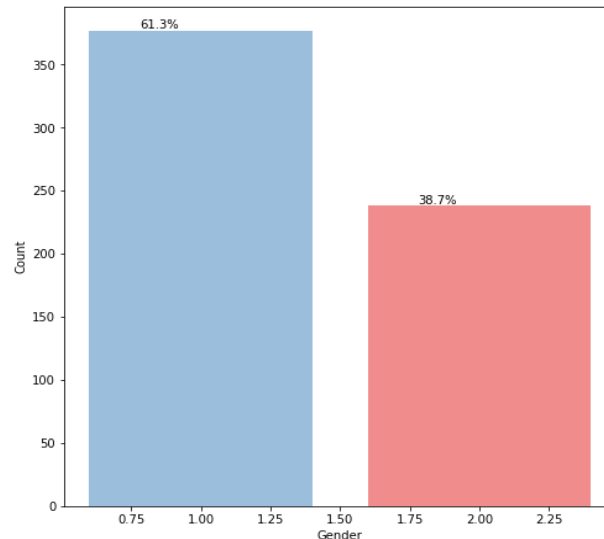


Figure 3. Sex of patients studied

Figure 4 shows the age distribution of the patients included in the data set. There is a notable concentration of individuals in the age range between 40 and 60 years, with a significant presence of patients aged 50 years. On the other hand, there is a smaller presence of patients in the 10 to 30 age range, as well as in those over 60 years of age. This disparity in age distribution underscores the importance of analyzing and understanding the demographic characteristics of the population under investigation.

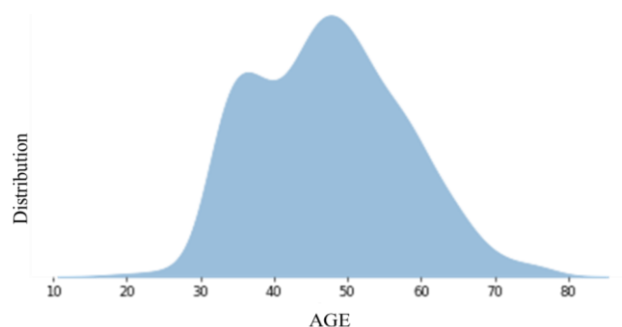


Figure 4. General distribution by age

Likewise, in Figure 5, a comparison was made between the age of the patients and their hepatic status. It is observed that those in the age range of 20 to 40 years present a higher propensity to develop hepatitis since there is a significantly higher concentration of cases in this interval in the data set analyzed. In addition, it is highlighted that patients who reach 50 years of age show a higher probability of developing fibrosis in the future. In a similar context, patients reaching 60 years of age show a higher probability of developing cirrhosis, while those between 40 and 50 years of age also exhibit a certain predisposition, although with a somewhat lower probability. However, it is important to note that there is a group of patients in good health in the 30-45 age range.

According to Figure 6, there is evidence of a greater propensity of men to develop liver disease compared to women. When examining Figure 6(a), it stands out that 5.3% of male patients present hepatitis, a figure that is equally significant in the case of cirrhosis, with a percentage of 5.3%, and 3.4% in fibrosis.

In contrast, in the female group, a lower incidence is observed, with only 1.7% affected by hepatitis, 4.2% by cirrhosis, and 3.4% by fibrosis as shown in Figure 6(b). These findings underscore the disparity in the prevalence of liver disease between men and women, suggesting a greater vulnerability of men to develop these types of conditions.

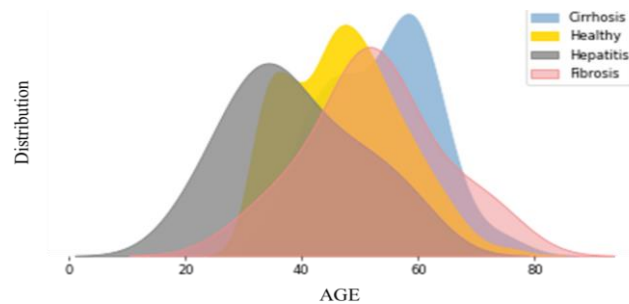


Figure 5. Distribution by age and liver status

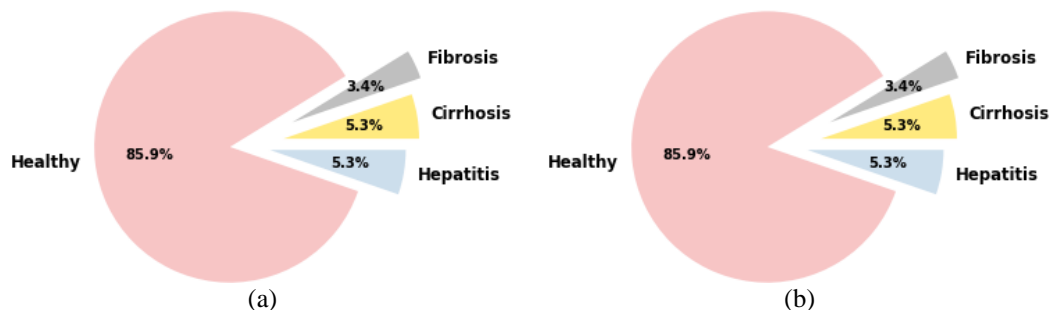


Figure 6. Distribution by sex and liver status: (a) males and liver status and (b) women and liver status

Likewise, according to Figure 7, Figure 7(a) a group of 533 individuals with a healthy liver was identified. However, it is important to highlight the presence of 24 patients diagnosed with hepatitis, 21 with hepatic fibrosis and 30 with cirrhosis. These liver health conditions demand specific attention and a comprehensive approach to ensure the well-being of those affected. In the case of hepatitis, Figure 7(b), it is recommended that patients receive constant medical follow-up, including laboratory tests to assess liver function and determine the effectiveness of treatment. Hepatic fibrosis, Figure 7(c), characterized by scar tissue formation in the liver, requires regular monitoring to assess disease progression. Patients are advised to take measures to mitigate risk factors, such as management of concurrent diseases and avoidance of hepatotoxic substances. In the case of cirrhosis, Figure 7(d), a more advanced and severe condition, it is critical to implement strategies to manage associated complications, such as ascites or hepatic encephalopathy. Patients with cirrhosis should strictly follow medical indications, including dietary sodium restriction and constant monitoring of liver function.

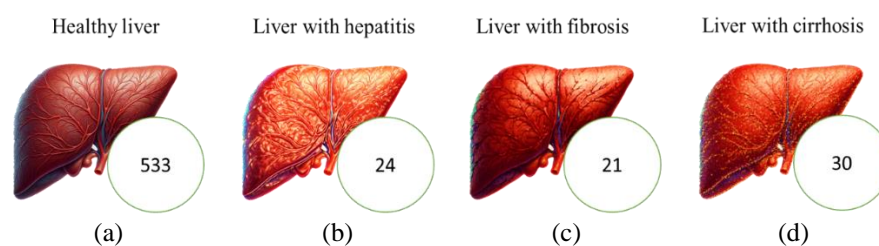


Figure 7. Liver status; (a) healthy liver, (b) liver with hepatitis, (c) liver with fibrosis, and (d) liver with cirrhosis

Figure 8 is target variable and blood tests. Figure 8(a) clearly shows that there is a direct correlation between the decrease in the amount of cholinesterase in the blood of patients and a significant increase in the probability of contracting hepatitis. This is most prominently manifested at the 7 and 9 g/dL levels of cholinesterase in the blood, where there is a notable concentration of confirmed cases. It is important to highlight that, in general terms, a mean of 7.5 g/dL of blood cholinesterase is evident in these cases. This finding reinforces the association between the low presence of cholinesterase and a predisposition to contract hepatitis, highlighting the relevance of monitoring and addressing the levels of this enzyme as a crucial factor in the prevention and diagnosis of the disease. Likewise, in Figure 8(b), we note a relationship between blood cholesterol levels and the probability of contracting hepatitis. This pattern reveals that as the amount of cholesterol in the blood decreases, the probability of contracting this disease increases. It is particularly noteworthy that a significantly higher concentration of cases in the 3 to 5 g/dL blood cholesterol range was identified. More specifically, the mean cholesterol in this range is observed to be 4.5 g/dL. These findings underline the importance of considering cholesterol levels as a relevant factor in the incidence of hepatitis.

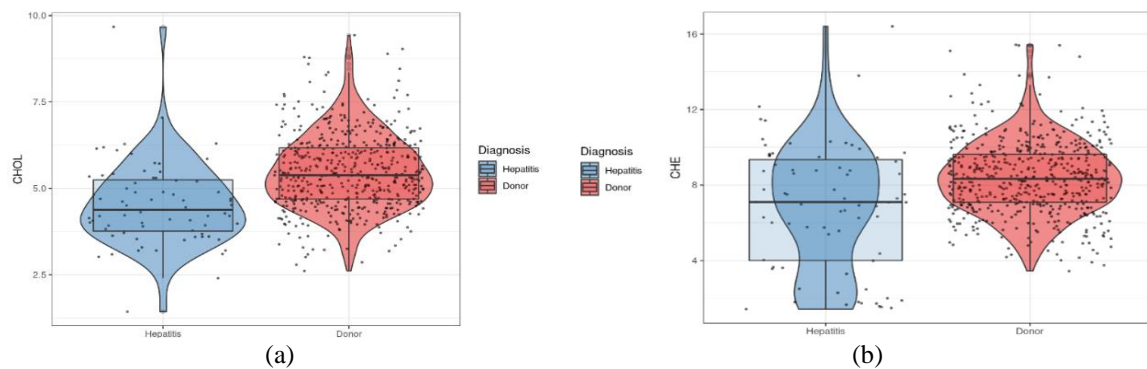


Figure 8. Target variable and blood tests; (a) target variable and amount of cholinesterase in the blood and (b) target variable and amount of cholesterol in the blood

In Figure 9, the impact of two additional characteristics of laboratory data on patient health is examined. In Figure 9(a), we contrast liver status with patients' age and blood cholinesterase concentration. We observe that as age increases and the amount of cholinesterase in the blood decreases, the likelihood of developing cirrhosis increases significantly. Conversely, when the presence of cholinesterase is higher but the patient's age is lower, the chances of contracting hepatitis increase significantly. Likewise, for ages between 20 and 70 years, and with an average cholinesterase level of 10, the odds of fibrosis increase. In Figure 9(b), it is highlighted that when the amount of albumin in the blood ranges between 40 and 50 g/dL, and the amount of alkaline phosphatase in the blood is low, the probabilities of hepatitis are higher, showing a behavior similar to that of fibrosis. On the other hand, when the blood albumin concentration is lower, the odds of cirrhosis are higher, regardless of the amount of alkaline phosphatase in the blood. As for Figure 9(c), it is observed that the lower the alanine transaminase concentration and the lower the amount of aspartate aminotransferase in the patient's blood, the greater the likelihood of developing cirrhosis. This pattern is similarly repeated in cases of hepatitis and fibrosis. These findings suggest a complex relationship between the variables analyzed and liver health, highlighting the importance of considering multiple factors to understand and prevent liver disease.

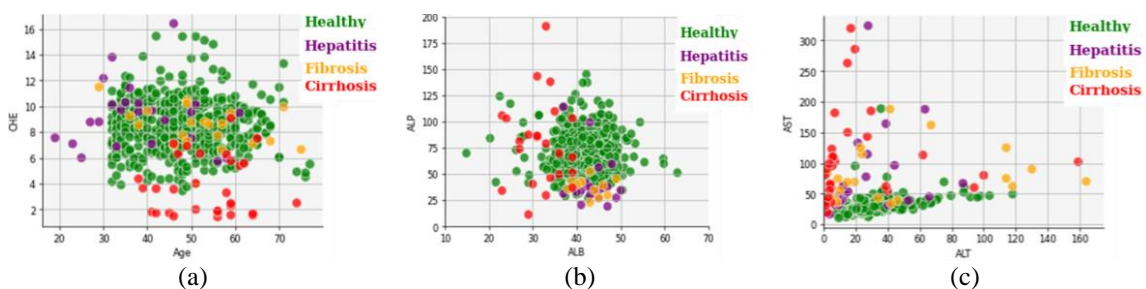


Figure 9. How liver health is affected by two characteristics: (a) age and acetylcholinesterase; (b) blood albumin and alkaline phosphatase; and (c) alanine transaminase and aspartate transaminase

3.2.4. Data preprocessing

After completing the data analysis, we divided the data set into two distinct groups. One portion was assigned for model evaluation, while the other portion was for model training. After splitting the data set, we proceeded to scale the data and train the corresponding models. This process ensures adequate performance evaluation and effective training of the models, thus contributing to the robustness and efficiency of the results. The importance of the clinical history as a fundamental variable and the technological intervention intrinsic to the modeling process made randomization impractical. Regarding blinding, we became aware of the internal complexity of ML models, so we chose to focus on transparency and reproducibility through detailed disclosure of the model architecture, hyperparameters, and evaluation methods. These methodological choices were based on the need to address the specific limitations of the study, with the aim of ensuring the integrity and ethics of the research.

4. RESULTS

Once the analysis and processing of the dataset were completed, we proceeded to train the ML models focused on hepatitis C prediction. The LR, DT, KNN, RF, and GB models were trained, to identify the model with the best performance in precision, accuracy, and sensitivity when predicting the disease. The results of these trainings are shown in Table 3.

The LR, RF, KNN, DT, and GB models achieved an accuracy of 89%, 93%, 85%, 95% and 94%, respectively. Likewise, in the accuracy indicator, the models registered 88%, 94%, 88%, 88%, 95% and 95%, respectively. The results highlight the DT model as the most effective predictor for hepatitis C, achieving a performance of 95% in accuracy, precision, sensitivity, and F1-score. It is closely followed by the GB model, with 94% in accuracy, 95% in precision, 94% in sensitivity, and 94% in F1-score. In third place is the RF model, with 93% in accuracy, 94% in precision, and 93% in sensitivity and F1-score. Despite not achieving metrics above 90%, the other models also obtained significant results. The KNN model achieved 85% in accuracy and sensitivity, 88% in precision, and 82% in F1-score. On the other hand, the LR model demonstrated solid performance with 89% accuracy, 88% in precision, 89% in sensitivity, and 87% in F1-score.

Table 3. Model training results

		Precision	Recall	F1-score	Support
Logistic regression	0	0.89	0.98	0.93	99
	1	0.86	0.50	0.63	24
	accuracy			0.89	123
	macro avg	0.87	0.74	0.78	123
	weighted avg	0.88	0.89	0.87	123
Random forest	0	0.93	0.99	0.96	99
	1	0.94	0.71	0.81	24
	accuracy			0.93	123
	macro avg	0.94	0.85	0.89	123
	weighted avg	0.94	0.93	0.93	123
KNN	0	0.85	1.00	0.92	99
	1	1.00	0.25	0.4	24
	accuracy			0.85	123
	macro avg	0.92	0.62	0.66	123
	weighted avg	0.88		0.85	0.82
Decision tree	0	0.95	0.99	0.97	99
	1	0.95	0.79	0.86	24
	accuracy			0.95	123
	macro avg	0.95	0.89	0.92	123
	weighted avg	0.95	0.95	0.95	123
Gradient boosting	0	0.93	1.00	0.97	99
	1	1.00	0.71	0.83	24
	accuracy			0.94	123
	macro avg	0.97	0.85	0.90	123
	weighted avg	0.95	0.94	0.94	123

5. DISCUSSION

HCV is a disease transmitted through contact with contaminated blood, claiming millions of lives annually. The most important findings of this study indicate that ML models can be effective tools for predicting hepatitis C. The LR, DT, KNN, RF, and GB models were evaluated. This evaluation was performed using performance metrics such as accuracy, precision, and sensitivity. Accuracy measured the

percentage of correct model predictions over total predictions, allowing us to assess how well it distinguishes between infected and uninfected patients. Accuracy indicated the reliability of the positive predictions, i.e., how many of the positive predictions are actually correct. Recall measured the proportion of true positives correctly identified by the model, which is crucial to ensure that the majority of hepatitis C cases are detected, minimizing false negatives and ensuring that few infected people go undetected.

The benchmarking process began with the collection and preparation of the dataset. In this study, a dataset with 615 records and 14 attributes was used, including patient demographic information and the results of clinical evaluations of blood and liver, among others. Subsequently, the LR, DT, KNN, RF and GB models were trained using this dataset. The training involved adjusting the model's parameters to minimize error in the predictions. After training, each model was evaluated for performance using a test dataset, calculating the accuracy, precision, and sensitivity metrics for each model. Subsequently, the results were compared, identifying the one with the best performance. After training, the models achieved the following results in terms of accuracy: LR (89%), RF (93%), KNN (85%), DT (95%), and GB (94%). In terms of accuracy, the results were: LR (88%), RF (94%), KNN (88%), DT (95%), and GB (95%). Of the results, the DT model showed the best overall performance with 95% in the metrics of accuracy, precision, sensitivity and F1-score. These results indicate that the DT model is the most suitable for predicting hepatitis C compared to the other models evaluated. DT's high accuracy and precision suggests that it can correctly identify both infected and uninfected patients, which is crucial for early detection and effective treatment of the disease.

This result is consistent with previous studies. For example, in the study from Kareem [30], the DT model achieved 93.44% accuracy using demographic data and clinical test results. However, in that study, they differed in data processing and optimizing ML models. Similarly, the GB model in our study achieved 94% accuracy, sensitivity, and F1-score, and 95% accuracy, compared to the study from Santos [26], where the model achieved 93.50% accuracy in predicting Hepatitis C. One of the coincidences with this study is the use of the same dataset, but differentiating with the application of 5-fold cross-validation. On the other hand, the RF model had a 93% performance in accuracy, sensitivity, and F1-score, results similar to those achieved in the studies [20], [23], [26], where the models achieved about 94% in accuracy, using additional techniques such as feature selection and forward sequential selection to improve their models, unlike our study where such techniques were not employed. The LR model in this study achieved a performance of 89% in accuracy and sensitivity, 88% in precision, and 87% in F1-score, similar to those recorded in [24], where the model had a performance of 82.9%, using the SMOTE oversampling technique to generate synthetic data and forward sequential selection to process the data, these being the main difference with our study. Finally, the KNN model was one of the last models with the lowest performance, with 85% accuracy and 88% precision. This is similar to the study from Ali *et al.* [24], where the model achieved 83% in accuracy, but differed significantly in studies such as [22], [27] where the model achieved 94.40% and 98.1% in accuracy, respectively, highlighting the use of optimization techniques and methods that were not employed in this study. Although the ML models evaluated achieved outstanding results consistent with previous studies, it is clear that the use of data optimization and processing techniques could further improve the performance of these models. Future studies should consider the integration of these techniques to maximize efficacy in predicting hepatitis C. It's important to also consider limitations, such as the size of the dataset and the variety of attributes. Larger, more diverse datasets could improve model generalizability.

The aim of this study was to benchmark different ML models for hepatitis C prediction, in order to determine which of the models offers better performance in terms of accuracy, precision, and sensitivity. The ability to predict hepatitis C accurately and early is crucial to improving detection and treatment rates of the disease. This study underscores the importance of implementing advanced predictive tools in the clinical setting to identify infected patients and administer appropriate treatments in a timely manner. This study contributes to the emerging field of digital health, demonstrating how ML models can be integrated into clinical practice to improve diagnostic accuracy and management of infectious diseases such as hepatitis C. The inclusion of other types of clinical data and biomarkers in future studies could further improve the accuracy and usefulness of predictive models.

6. CONCLUSION

HCV infection is a disease with no cure available today, affecting millions of people of all ages around the world. It spreads mainly through contact with contaminated blood, through injections, transfusions, and other means. Given that up to 70% of infected individuals can achieve a successful recovery if they receive treatment in a timely manner, it is crucial to develop techniques that make it easier for medical professionals to detect this pathology early. In this study, five ML models focused on the prediction of Hepatitis C were developed, analyzed, and evaluated, with the aim of determining which of the models offers the best performance in this task. After analyzing, processing, and training the models, the results showed

that the DT model achieved the best performance in terms of accuracy, precision, sensitivity and F1-score, reaching 95%. The GB and RF models also demonstrated good performance, with an accuracy of 94% and 93%, respectively. It is important to note that during the univariate and multivariate analysis stage of the dataset, we observed that certain indicators allow us to predict with high probability whether a person suffers from hepatitis C. Factors such as age and sex were found to be determinants, as older people and men are more likely to contract hepatitis C. In addition, indicators such as cholesterol and cholinesterase levels in the blood were also significant: the lower these levels, the higher the likelihood of HCV infection.

Although the results of this study are promising, it is essential to consider that optimization techniques play a crucial role in the accuracy of ML models. A poor optimization or data processing process can negatively affect training results. Therefore, it is recommended that appropriate techniques and methods be employed in future studies to improve the effectiveness of these predictive models further. Specifically, future research should explore the integration of more advanced optimization techniques and the use of broader and more varied datasets to improve the accuracy and generalizability of predictive models. In addition, the inclusion of new types of clinical data and biomarkers could provide valuable additional information for predicting hepatitis C. It is essential to continue to develop and refine advanced predictive tools to facilitate the early detection and effective treatment of hepatitis C, which would contribute to improving recovery rates and lessening the impact of this disease globally.

FUNDING INFORMATION

This research received no external funding.

AUTHOR CONTRIBUTIONS STATEMENT

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Michael Cabanillas-Carbonell	✓	✓		✓	✓			✓	✓		✓	✓		
Joselyn Zapata-Paulini				✓		✓		✓	✓	✓	✓			✓

C : **C**onceptualization

M : **M**ethodology

So : **S**oftware

Va : **V**alidation

Fo : **F**ormal analysis

I : **I**nvestigation

R : **R**esources

D : **D**ata Curation

O : **O**riginal Draft

E : **E**diting

Vi : **V**isualization

Su : **S**upervision

P : **P**roject administration

Fu : **F**unding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

INFORMED CONSENT

Not applicable. This study did not involve direct participation of human subjects or animals. The analysis was conducted using publicly available and anonymized datasets, which are freely accessible for research purposes. Therefore, individual informed consent was not required.

ETHICAL APPROVAL

Not applicable-No direct intervention with human subjects or animals.

DATA AVAILABILITY

Data will be made available on request.

REFERENCES




- [1] R. H. Westbrook and G. Dusheiko, "Natural history of hepatitis C," *Journal of Hepatology*, vol. 61, no. 1, pp. S58–S68, Nov. 2014, doi: 10.1016/j.jhep.2014.07.012.
- [2] World Health Organization, "Hepatitis C," World Health Organization Accessed: Dec. 19, 2023. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/hepatitis-c>.
- [3] T. Pietschmann and R. J. P. Brown, "Hepatitis C virus," *Trends in Microbiology*, vol. 27, no. 4, pp. 379–380, Apr. 2019, doi: 10.1016/j.tim.2019.01.001.

- [4] C. W. Shepard, L. Finelli, and M. J. Alter, "Global epidemiology of hepatitis C virus infection," *Lancet Infectious Diseases*, vol. 5, no. 9, pp. 558–567, Sep. 2005, doi: 10.1016/S1473-3099(05)70216-4.
- [5] K. M. Hanafiah, J. Groeger, A. D. Flaxman, and S. T. Wiersma, "Global epidemiology of hepatitis C virus infection: new estimates of age-specific antibody to HCV seroprevalence," *Hepatology*, vol. 57, no. 4, pp. 1333–1342, Apr. 2013, doi: 10.1002/hep.26141.
- [6] D. Lavanchy, "Evolving epidemiology of hepatitis C virus," *Clinical Microbiology and Infection*, vol. 17, no. 2, pp. 107–115, Feb. 2011, doi: 10.1111/j.1469-0691.2010.03432.x.
- [7] A. A. Mohamed, T. A. Elbedewy, M. El-Serafy, N. El-Toukhy, W. Ahmed, and Z. A. El Din, "Hepatitis C virus: a global view," *World Journal of Hepatology*, vol. 7, no. 26, pp. 2676–2680, 2015, doi: 10.4254/wjh.v7.i26.2676.
- [8] J. P. Messina *et al.*, "Global distribution and prevalence of hepatitis C virus genotypes," *Hepatology*, vol. 61, no. 1, pp. 77–87, Jan. 2015, doi: 10.1002/hep.27259.
- [9] B. Hajarizadeh, J. Grebely, and G. J. Dore, "Epidemiology and natural history of HCV infection," *Nature Reviews Gastroenterology and Hepatology*, vol. 10, no. 9, pp. 553–562, Sep. 2013, doi: 10.1038/nrgastro.2013.107.
- [10] D. G. Murphy, B. Willems, M. Deschênes, N. Hilzenrat, R. Mousseau, and S. Sabbah, "Use of sequence analysis of the NS5B region for routine genotyping of hepatitis C virus with reference to C/E1 and 5' untranslated region sequences," *Journal of Clinical Microbiology*, vol. 45, no. 4, pp. 1102–1112, Apr. 2007, doi: 10.1128/JCM.02366-06.
- [11] S. D. Holmberg, P. R. Spradling, A. C. Moorman, and M. M. Denniston, "Hepatitis C in the United States," *New England Journal of Medicine*, vol. 368, no. 20, pp. 1859–1861, May 2013, doi: 10.1056/nejmp1302973.
- [12] M. M. Denniston, R. M. Kleven, G. M. McQuillan, and R. B. Jiles, "Awareness of infection, knowledge of hepatitis C, and medical follow-up among individuals testing positive for hepatitis C: National Health and Nutrition Examination Survey 2001–2008," *Hepatology*, vol. 55, no. 6, pp. 1652–1661, Jun. 2012, doi: 10.1002/hep.25556.
- [13] World Health Organization, "Viral hepatitis report by the secretariat," *World Health Organization*, 2010.
- [14] Y. Baashar *et al.*, "Effectiveness of artificial intelligence models for cardiovascular disease prediction: network meta-analysis," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–12, Feb. 2022, doi: 10.1155/2022/5849995.
- [15] O. Iparraquirre-Villanueva *et al.*, "Comparison of predictive machine learning models to predict the level of adaptability of students in online education," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 4, pp. 494–503, 2023, doi: 10.14569/IJACSA.2023.0140455.
- [16] V. R. Karna and K. V. V. Reddy, "1-dimensional convolutional neural networks for predicting sudden cardiac," *IAES International Journal of Artificial Intelligence(IJ-AI)*, vol. 13, no. 1, pp. 984–993, Mar. 2024, doi: 10.11591/ijai.v13.i1.pp984-993.
- [17] M. Cabanillas-Carbonell and J. Zapata-Paulini, "Evaluation of machine learning models for the prediction of Alzheimer's: In search of the best performance," *Brain Behav Immun Health*, vol. 44, p. 100957, 2025, doi: 10.1016/j.bbih.2025.100957.
- [18] N. Absar *et al.*, "The efficacy of machine-learning-supported smart system for heart disease prediction," *Healthcare (Switzerland)*, vol. 10, no. 6, p. 1137, Jun. 2022, doi: 10.3390/healthcare10061137.
- [19] A. Alizargar, Y. L. Chang, and T. H. Tan, "Performance comparison of machine learning approaches on hepatitis c prediction employing data mining techniques," *Bioengineering*, vol. 10, no. 4, 2023, doi: 10.3390/bioengineering10040481.
- [20] L. Syafaah, Z. Zulfatman, I. Pakaya, and M. Lestandy, "Comparison of machine learning classification methods in hepatitis C virus," *Jurnal Online Informatika*, vol. 6, no. 1, pp. 73–78, Jun. 2021, doi: 10.15575/join.v6i1.719.
- [21] L. Ma, Y. Yang, X. Ge, Y. Wan, and X. Sang, "Prediction of disease progression of chronic hepatitis C based on XGBoost algorithm," in *Proceedings - 2020 International Conference on Robots and Intelligent Systems, ICRIS 2020*, Nov. 2020, pp. 598–601, doi: 10.1109/ICRIS52159.2020.00151.
- [22] K. Ahammed, M. S. Satu, M. I. Khan, and M. Whaiduzzaman, "Predicting infectious state of hepatitis C virus affected patient's applying machine learning methods," in *2020 IEEE Region 10 Symposium, TENSYP 2020*, 2020, pp. 1371–1374, doi: 10.1109/TENSYP50017.2020.9230464.
- [23] H. M. Farghaly, M. Y. Shams, and T. Abd El-Hafeez, "Hepatitis C virus prediction based on machine learning framework: a real-world case study in Egypt," *Knowledge and Information Systems*, vol. 65, no. 6, pp. 2595–2617, Jun. 2023, doi: 10.1007/s10115-023-01851-4.
- [24] A. M. Ali *et al.*, "Explainable machine learning approach for hepatitis C diagnosis using SFS feature selection," *Machines*, vol. 11, no. 3, p. 391, Mar. 2023, doi: 10.3390/machines11030391.
- [25] L. Chen, P. Ji, and Y. Ma, "Machine learning model for hepatitis C diagnosis customized to each patient," *IEEE Access*, vol. 10, pp. 106655–106672, 2022, doi: 10.1109/ACCESS.2022.3210347.
- [26] D. Santos, "Predicting the severity of hepatitis C using machine learning models." Oct. 16, 2023, doi: 10.20944/preprints202310.0952.v1.
- [27] V. Harabor *et al.*, "Machine learning approaches for the prediction of hepatitis B and C seropositivity," *International Journal of Environmental Research and Public Health*, vol. 20, no. 3, p. 2380, Jan. 2023, doi: 10.3390/ijerph20032380.
- [28] S. M. A. El-Salam *et al.*, "Performance of machine learning approaches on prediction of esophageal varices for Egyptian chronic hepatitis C patients," *Informatics in Medicine Unlocked*, vol. 17, p. 100267, 2019, doi: 10.1016/j.imu.2019.100267.
- [29] S. Hashem *et al.*, "Comparison of machine learning approaches for prediction of advanced liver fibrosis in chronic hepatitis C patients," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 15, no. 3, pp. 861–868, May 2018, doi: 10.1109/TCBB.2017.2690848.
- [30] L. S. W. T. A. Kareem, "Development of diagnostic decision making for chronic hepatitis C virus patients by various supervised predictive model," *Journal of Advanced Research in Dynamic and Control Systems*, vol. Volume 12, no. Issue 6, pp. 3113–3123, 2020, doi: 10.5373/JARDCS/V12I6/S20201276.
- [31] U. K. Lilhore *et al.*, "Hybrid model for precise hepatitis-C classification using improved random forest and SVM method," *Scientific Reports*, vol. 13, no. 1, p. 12473, Aug. 2023, doi: 10.1038/s41598-023-36605-3.
- [32] T. M. Ghazal *et al.*, "Hep-pred: Hepatitis C staging prediction using fine gaussian SVM," *Computers, Materials and Continua*, vol. 69, no. 1, pp. 191–203, 2021, doi: 10.32604/cmc.2021.015436.
- [33] M. Saw, T. Saxena, S. Kaithwas, R. Yadav, and N. Lal, "Estimation of prediction for getting heart disease using logistic regression model of machine learning," in *2020 International Conference on Computer Communication and Informatics, ICCCI 2020*, Jan. 2020, pp. 1–6, doi: 10.1109/ICCCI48352.2020.9104210.
- [34] H. K. Andi, "An accurate bitcoin price prediction using logistic regression with LSTM machine learning model," *Journal of Soft Computing Paradigm*, vol. 3, no. 3, pp. 205–217, Sep. 2021, doi: 10.36548/jscp.2021.3.006.
- [35] H. C. Lee *et al.*, "Prediction of acute kidney injury after liver transplantation: machine learning approaches vs. logistic regression model," *Journal of Clinical Medicine*, vol. 7, no. 11, p. 428, Nov. 2018, doi: 10.3390/jcm7110428.




- [36] F. Yu, C. Wei, P. Deng, T. Peng, and X. Hu, "Deep exploration of random forest model boosts the interpretability of machine learning studies of complicated immune responses and lung burden of nanoparticles," *Science Advances*, vol. 7, no. 22, May 2021, doi: 10.1126/sciadv.abf4130.
- [37] J. Song *et al.*, "The random forest model has the best accuracy among the four pressure ulcer prediction models using machine learning algorithms," *Risk Management and Healthcare Policy*, vol. 14, pp. 1175–1187, Mar. 2021, doi: 10.2147/RMHP.S297838.
- [38] J. Zapata-Paulini and M. Cabanillas-Carbonell, "Performance analysis of 10 machine learning models in lung cancer prediction," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 37, no. 2, p. 1352, 2025, doi: 10.11591/ijeecs.v37.i2.pp1352-1364.
- [39] N. Peppes, E. Daskalakis, T. Alexakis, E. Adamopoulou, and K. Demestichas, "Performance of machine learning-based multi-model voting ensemble methods for network threat detection in agriculture 4.0," *Sensors*, vol. 21, no. 22, p. 7475, Nov. 2021, doi: 10.3390/s21227475.
- [40] D. Prasad, S. K. Goyal, A. Sharma, A. Bindal, and V. S. Kushwah, "System model for prediction analytics using /(-nearest neighbors algorithm," *Journal of Computational and Theoretical Nanoscience*, vol. 16, no. 10, pp. 4425–4430, Oct. 2019, doi: 10.1166/jctn.2019.8536.
- [41] F. Nigsch, A. Bender, B. Van Buuren, J. Tissen, E. Nigsch, and J. B. O. Mitchell, "Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization," *Journal of Chemical Information and Modeling*, vol. 46, no. 6, pp. 2412–2422, Nov. 2006, doi: 10.1021/ci060149f.
- [42] C. Kingsford and S. L. Salzberg, "What are decision trees?," *Nature Biotechnology*, vol. 26, no. 9, pp. 1011–1012, Sep. 2008, doi: 10.1038/nbt0908-1011.
- [43] B. de Ville, "Decision trees," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 5, no. 6, pp. 448–455, Nov. 2013, doi: 10.1002/wics.1278.
- [44] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in Neurobotics*, vol. 7, no. DEC, 2013, doi: 10.3389/fnbot.2013.00021.
- [45] A. J. Ferreira and M. A. T. Figueiredo, "Boosting algorithms: a review of methods, theory, and applications," in *Ensemble Machine Learning*, New York, NY: Springer New York, 2012, pp. 35–85.
- [46] A. Mayr, H. Binder, O. Gefeller, and M. Schmid, "The evolution of boosting algorithms," *Methods of Information in Medicine*, vol. 53, no. 06, pp. 419–427, Jan. 2014, doi: 10.3414/ME13-01-0122.
- [47] P. Bühlmann and B. Yu, "Boosting," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 1, pp. 69–74, Jan. 2010, doi: 10.1002/wics.55.

BIOGRAPHIES OF AUTHORS



Michael Cabanillas-Carbonell    is engineer and master in systems engineering from the National University of Callao-Peru, Ph.D. candidate in systems engineering and telecommunications at the Polytechnic University of Madrid. Ex president of the chapter of the Education Society IEEE-Peru. Conference chair of the Engineering International Research Conference IEEE Peru EIRCON. Specialization in software development, artificial intelligence, machine learning, business intelligence, augmented reality. Reviewer IEEE Peru and author of more than 100 scientific articles indexed in IEEE Xplore and Scopus. He can be contacted at email: mcabanillas@ieee.org.



Joselyn Zapata-Paulini    is bachelor in systems engineering and computer science from the Universidad de Ciencias y Humanidades, master in science with environmental management and sustainable development at the Universidad Continental, Peru. She has several international publications. Specialized in the areas of augmented reality, virtual reality, and the internet of things. Author of scientific articles indexed in IEEE Xplore, Scopus, and WoS. She can be contacted at email: 70994337@continental.edu.pe.