Predicting hepatitis C infection with machine learning algorithms: a prospective study

Orlando Iparraguirre-Villanueva[1], Rosalynn Ornella Flores-Castañeda[2], Henry Chero-Valdivieso[2], Fernando Sierra-Liñan[2]
[1]Facultad de Ingeniería, Universidad Tecnológica del Perú, Lima, Perú
[2]Facultad de Ingeniería y Arquitectura, Universidad César Vallejo, Lima, Perú

Article Info

Article history:

Received Nov 30, 2023
Revised Mar 25, 2024
Accepted Apr 17, 2024

Keywords:

Hepatitis C
Infection
Machine learning
Models
Prediction

ABSTRACT

Globally, chronic hepatitis C virus (HCV) infection affects millions of people and leads to a high number of deaths annually. In 2019, the World Health Organization (WHO) recorded around 290,000 deaths related to HCV, a virus transmitted mainly through blood that causes liver damage. The virus has infected more than 169 million people worldwide. This study aims to compare the performance of machine learning (ML) models for HCV detection. ML models such as logistic regression (LR), random forest (RF), decision tree (DT), and catBoost classifier (CATBC) were used. To carry out this task, a dataset of 615 patient records, and 14 variables were used. This research process was carried out in multiple phases, encompassing model understanding, data analysis and cleaning, ML model training, and subsequent model evaluation. The results revealed that the gradient boosting (GB) model stood out by achieving the best performance and highest accuracy, achieving a rate of 94% in HCV detection, this demonstrates outstanding performance compared to the other models such as LR, RF, k-nearest neighbor (KNN), DT, and CATBC, which obtained accuracy rates of 89%, 93%, 85%, 93%, 93%, and 92%, respectively. It can be concluded that the GB model stands out as the best algorithm for this task.

Corresponding Author:

Orlando Iparraguirre-Villanueva
Facultad de Ingeniería, Universidad Tecnológica del Perú, Lima, Perú
Email: oiparraguirre@ieee.org

1. INTRODUCTION

Globally, an estimated 58 million people are living with chronic hepatitis C virus (HCV) infection [1]. In 2019, the WHO reported that approximately 290,000 people lost their lives to HCV [2]. This virus, which spreads through the blood and causes liver damage [3], has infected more than 169 million people worldwide [4]. In terms of the impact of HCV, around 20% of patients experience an acute form of hepatitis, while 75% to 85% of those affected develop chronic health conditions [5]. Specifically, types B and C of this virus are notorious for inducing chronic diseases, such as liver cirrhosis and cancer [6]. Over time, chronic HCV infection can have serious consequences, including the development of end-stage liver disease and hepatocellular carcinoma [7]. Despite the pressing need for a preventive solution, no effective vaccine against HCV infection has been developed so far [8]. In addition, about 70% of HCV infected patients experience chronic disease, while the remainder undergo acute and transient infection. In addition, about 70% of HCV infected patients experience chronic disease, while the remainder experience acute and transient infection [9]. Overall, HCV presents as a global health problem, leading to the development of liver cancer, and especially affecting marginalized people with limited access to traditional health services, including testing and treatment [10], [11].

There are places where the availability of HCV testing and treatment is insufficient [12]. In addition, it is noted that chronic HCV infection in children is generally symptomless or mildly symptomatic. However, over time, end-stage liver disease requiring liver transplantation can progress to substantial fibrosis, cirrhosis, hepatocellular carcinoma, and other conditions [13]. Particularly in Puerto Rico, people who inject drugs are disproportionately affected by HCV amid an increase in HIV and HCV infections in people who use drugs [14], [15]. Although hepatitis C is curable, only 21% of people with HCV infection are diagnosed and only 13% have received curative treatment [16]. This infection usually occurs through blood transfusions or the sharing of shaving tools and can even occur through sexual practices [17]. Importantly, HCV infection is more common in people living with HIV [18]. In the city of Montreal, Canada, a high incidence of HCV persists, with 21 cases per 100 people per year in 2017, especially among people who inject drugs [19]. HCV causes liver inflammation and leads to acute and chronic hepatitis [20]. Chronic HCV infection imposes considerable health and economic burdens on patients and society at large [21]. Although screening is the first step in the HCV continuum of care, it is still unclear how to improve it, especially in hard-to-reach populations [22]. This unpredictable disease can worsen the human health situation if not properly diagnosed [23].

Artificial intelligence (AI) applications have seen a significant increase in their use in medical and healthcare settings in the last five years [24]. Machine learning (ML) is noted for its effectiveness in providing accurate and precise information for the diagnosis of various diseases [25]. Traditionally, ML has been used in medical practice to aid in patient diagnosis through deep learning and medical image analysis [26]. With the continuous advancement of information technology and the growth of medical data, more and more medical professionals are recognizing the potential of AI, and some even believe that this technology could completely transform medical practice using advanced ML methods [27]. ML applications are revolutionizing medicine [28], especially given the extensive use of this technology in predicting patient outcomes [29]. In the medical field, where misdiagnosis can have serious consequences, supervised ML techniques have demonstrated their potential to outperform conventional diagnostic methods, thus helping medical professionals to identify high-risk diseases more accurately [30].

In recent years researchers and academics have written articles related to the topic of study. For example, Ma *et al.* [31] aimed to diagnose early progression of chronic HCV in patients with this disease. For this, they used the XGBoost model, support vector machine (SVM), k-nearest neighbor (KNN), decision tree (DT), and AdaBoost, achieving the highest accuracy of 91.56% with XGBoost. Also, Islam *et al.* [32] used ML models to predict HCV, for which they worked with naive Bayes (NB), random forest (RF), KN, DT, deep learning, and artificial neural network (ANN) algorithms. After using various algorithms, ANN shows the best results, with an accuracy of 95.50%. Similarly, Hafeez *et al.* [33] studied different algorithms for diagnosing liver disease, comparing linear regression, DT, RF, KNN, and SVM. The results showed that SVM obtained the best metrics with an accuracy of 91.84%. Rouhani and Haghighi [34] diagnosed hepatitis using SVM and ANN, achieving 97% prediction. Also, Olatunji *et al.* [35] performed a comparative analysis of different ML models for HCV prevention, using DT, KNN, and NB. The results showed that KNN achieved the best metrics with 86.05% accuracy. On the other hand, Saputra *et al.* [36] proposed RF for HCV classification, they compared this algorithm with NB, KNN, and DT. RF obtained the best accuracy metrics with 99.46%. Similarly, Singh *et al.* [37] developed a highly optimized XGBoost algorithm for the anticipation of early progression to hepatitis C. The designed methodology produced predictions of early hepatitis C progression. The designed methodology produced HCV progression predictions with an exceptional accuracy of 98.6%, significantly outperforming other algorithms such as logistic regression (LR) LightGBM (LGBM), DT, and SVM-radial basis function (RBF). Shahzadi *et al.* [38] sought to predict HCV by ML algorithm using KNN, DT, support vector classifier (SVC), and multilayer perceptron (MLP). MLP achieved the best metrics with 95.9% accuracy. Trishna *et al.* [39] analyzed different techniques for hepatitis A, B, and C detection. They used ML models such as NB, KNN, and RF. Using cross-validation, the RF algorithm achieved an accuracy of 98.6%.

This research aims to compare the performance of ML models in HCV detection using ML algorithms such as LR, RF, KNN, DT, CatBoost (CB), and gradient boosting classifier (GBC). The article follows a structure composed of six sections. In the first section, a context is provided for the problem addressed in the case study. In the second section, reference is made to articles related to the central theme of the article. The third section is devoted to a detailed description of the methodology used. Then, in the fourth section, the results obtained are presented. Finally, in the last two sections, the results are analyzed and discussed, and the conclusions drawn from the research are presented.

2. METHOD

In this section of the study, the theoretical foundations underlying the ML models, such as LR, RF, KNN, DT, CB, and GBC, are provided, along with an explanation of the approach used in HCV prediction. In addition, the development of the case study is presented.

2.1. Logistic regression

LR is a classification method that is based on probabilities and has performed remarkably well in several areas, even being applied in situations involving multiple instances [40]. Its main use lies in binary classification [41], where the relationship between binary dependent variables (output variables) and explanatory variables (input variables) is modeled using a probabilistic statistical approach [42]. Primarily designed to solve two-class classification problems, this algorithm estimates the probability of an event occurring, and its output corresponds to the probability that the model predicts that the test samples belong to the positive class [43]. In (1) represents the LR model.

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

In this context, Y symbolizes the likelihood of an event occurring, which is represented as P(Y).

2.2. Random forest

RF is an improved classifier based on the construction of multiple DT. During the creation of these trees, variables are assigned importance to determine their relevance in the process [44] RF operates by selecting random samples with replacement and building DT without pruning at each iteration. The contribution of all trees is then combined through a "voting" process to determine the most popular class, resulting in a robust prediction [45]. DT, which are an integral part of the RF, are noted for their efficiency in classifying data based on their most distinctive features [46]. Figure 1 presents the RF architecture.

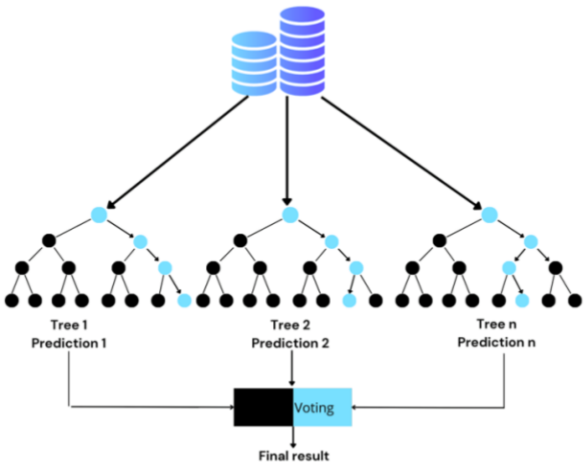


Figure 1. RF architecture

2.3. K-nearest neighbors

KNN is often considered a "delayed learning" or "memory-based" approach because it generates predictions for new cases by considering the k closest or similar training examples, without the need to build a model during a dedicated training phase [47]. Being one of the most widely employed classifiers in the ML field, KNN finds application in solving reliability-related problems [48]. Moreover, it stands out for its simplicity and fundamental character as a non-parametric local approximation method, suitable for both classification and regression tasks [49]. In a specific classification task, KNN classifies an unlabeled test sample by a majority vote based on the KNN belonging to all classes, selected by a specific distance or dissimilarity metric depending on the types of attributes involved [50]. The Euclidean formula used in this model is presented in (2).

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

Within this expression, the symbols "*x*" and "*y*" represent vectors indicating two examples within the feature space, while "*xi*" and "*yi*" refer to the individual components of the vector's "*x*" and "*y*", respectively. In addition, "*n*" corresponds to the number of attributes present in the feature space.

2.4. Decision tree classifier

DT represents an ML approach used in the evaluation of forecasts [51]. This model can handle non-linear relationships and capture interactions between variables, thus improving its accuracy in predicting outcomes [52]. DTs work by recursively dividing data into smaller groups, selecting optimal features based on parameters such as information gain or the Gini index [53]. This supervised tree-based algorithm predicts numerical results by identifying local regions through recursive partitioning in fewer steps. It is composed of three types of nodes: the root node, interior nodes, and leaf nodes, and is based on decision criteria that reflect the characteristics of the data collection [54]. In (3), the DT model is represented.

$$E(s) = \sum_{k=0}^{n} \binom{n}{k} - Py * \log 2Pn \tag{3}$$

Within this formula, "*E*" symbolizes the quantification of the degree of disorder or uncertainty, while "s" represents the sample. In addition, "*Py*" is used to indicate the probability that the event in question will occur, while "*Pn*" denotes the probability that the event will not materialize.

2.5. CatBoost classifier

The CB algorithm excels in its efficient handling of categorical features during training, avoiding overfitting thanks to its unbiased gradient estimation [55]. This leads to a significant reduction in dependence on a wide variety of hyperparameter settings [56]. In addition, CB implements an effective strategy that decreases the risk of overfitting, allowing full utilization of the training dataset [57]. This gradient-boosting DT-based ML framework differentiates itself by creating new trees by adapting to the gradient of the current model, overcoming the gradient bias problems common in traditional gradient-boosting algorithms [58]. CB is depicted in (4).

$$\mathcal{L}(H) \coloneqq \mathbb{E}\mathrm{L}(\mathcal{y}, H(\mathrm{X})) \tag{4}$$

Within this context, the smooth loss function is denoted as L (.,.) and the pair (X, y) refers to a test instance that has been obtained by a sampling process from the training data set.

2.6. Gradient boosting classifier

Gradient boosting (GB) is a fundamental strategy in ML, which consists of combining weak predictors into a stronger predictor and is especially useful in classification, regression, and other domains [59]. This ensemble learning approach differs from the traditional method, as it assembles a set of weak models to build a stronger and more effective model [60]. The learning process of a gradient boosting machine is based on iterative model improvement based on the residuals between the predictions generated by previous models and the true values [61]. In (5) expresses the mathematical equation of the model.

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

In this description, $\hat{y}$ refers to the final accuracy of the model, f(x) denotes the prediction function, γ represents the learning coefficient, and *h(x)* corresponds to the prediction produced by the least robust model at the i-th iteration.

2.7. Understanding the dataset

The dataset used in this research was obtained from the Kaggle platform, consisting of a total of 615 records, and 14 attributes. These attributes include: "ID", which represents a unique identification number for each patient; "category", a variable coding for different health conditions (blood donor, suspected blood donor, hepatitis, for fibrosis, and cirrhosis); "age" and "gender" for the patient's age and gender respectively; "albumin (ALB)", indicating abnormal blood albumin levels, related to liver function and blood proteins; "alkaline phosphatase (ALP)", an enzyme measured in blood tests that has implications for bone and liver health; "alanine aminotransferase (ALT)", used as a primary indicator of liver health; "aspartate aminotransferase (AST), also a liver indicator; bilirubin (BIL), which measures the concentration of this substance in the blood and is related to liver function; cholinesterase (CHE), an enzyme measured in blood tests related to neuromuscular function; "cholesterol (CHOL)", which measures the total cholesterol level, relevant to cardiovascular health; "creatinine (CREA)", which assesses kidney function; and "gamma-glutamyl transferase (GGT)", which measures this enzyme in blood serum, being indicative of liver and biliary health; finally, "protein (PROT)" refers to the total protein level in the blood. Figure 2 shows a graphic representation of the development process of this research.

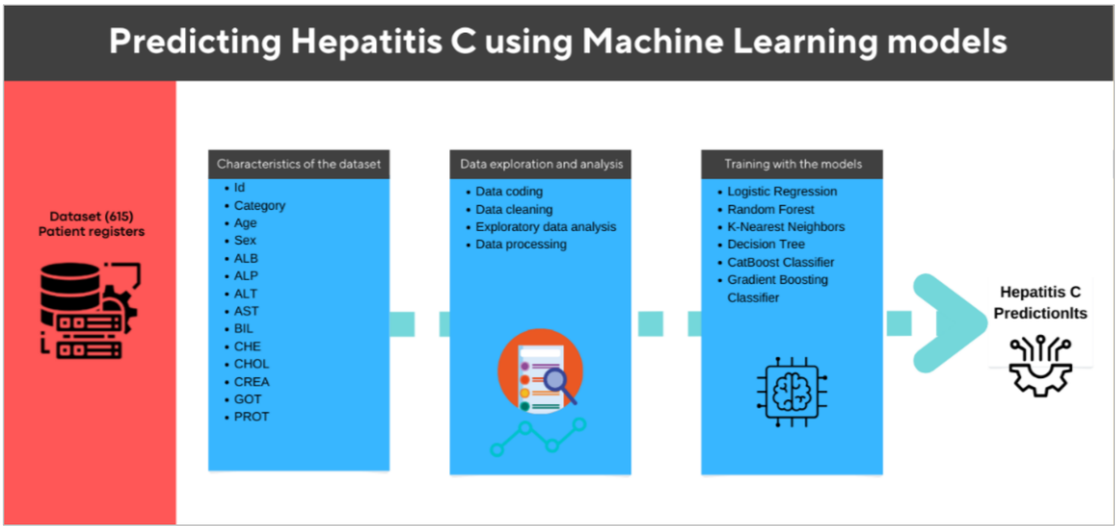


Figure 2. Case study development process

2.7.1. Understanding the dataset

In this section, an initial assessment of the content of the dataset is carried out before proceeding with the analysis and training of the ML models as evidenced in Table 1. At the beginning of the process, imports of essential libraries for data manipulation were performed, including seaborn, matplotlib, NumPy, and Pandas. In addition, a verification of the data types in each variable was carried out to ensure the most effective training approach. During this analysis phase, a transformation of the variable "Sex" was performed, where "m" was replaced by 1 and "f" by 2 to represent male and female genders, respectively. In addition, adjustments were made to the data for the variable "category", replacing "0= blood donor" and "0s= suspect blood donor" with 0, and modifying "1=hepatitis", "2=fibrosis" and "3: cirrhosis" to 1, to simplify and standardize the categorization. Also, the statistics of the dataset were analyzed as shown in Table 2, these statistics are fundamental to understanding the distribution and characteristics of the variables in the dataset. For example, the average age is approximately 47 years, with a variability of about 10 years, and the average ALB and ALP concentrations are approximately 41.6 and 68.3 respectively. On the other hand, ALT levels on average are around 28.45, but there is a lot of variability in these levels, as the standard deviation is high at 25.47. The lowest value recorded is 0.9, indicating that some people have very low levels of ALT in their blood. In the case of CHOL, the mean is 5.37, with a standard deviation of 1.13. The minimum value is 1.43, which tells us that most of the observations have CHOL values close to the mean. This descriptive data is essential to identify trends, outliers, and patterns in the data.

Table 1. Characteristics of the dataset

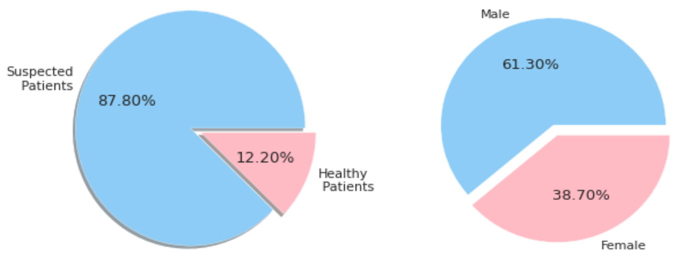
#	Category	Age	Sex	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT
1	0=Blood Donor	32	m	38.5	52.5	7.7	22.1	7.5	6.93	3.23	106	12.1	69
2	0=Blood Donor	32	m	38.5	70.3	18	24.7	3.9	11.17	4.8	74	15.6	76.5
3	0=Blood Donor	32	m	46.9	74.7	36.2	52.6	6.1	8.84	5.2	86	33.2	79.3
4	0=Blood Donor	32	m	43.2	52	30.6	22.6	18.9	7.33	4.74	80	33.8	75.7
5	0=Blood Donor	32	m	39.2	74.1	32.6	24.8	9.6	9.15	4.32	76	29.9	68.7

Table 2. Descriptive statistics of the variables in the dataset

	Count	Mean	Std	Min	25%	50%	75%	Max
Age	615	47.40813	10.055105	19	39	47	54	77
ALB	614	41.620	5.781	14.9	38.8	41.95	45.2	82.2
ALP	597	68.283	26.028	11.3	52.5	66.2	80.1	416.6
ALT	614	28.450	25.469	0.9	16.4	23	33.075	325.3
AST	615	34.786	33.090	10.6	21.6	25.9	32.9	324
BIL	615	11.396	19.673	0.8	5.3	7.3	11.2	254
CHE	615	8.196	2.2056	1.42	6.935	8.26	9.59	16.41
CHOL	605	5.368	1.133	1.43	4.61	5.3	6.06	9.67
CREA	615	81.287	49.756	8	67	77	88	1079.1
GGT	615	39.533	54.661	4.5	15.7	23.3	40.2	650.9
PROT	614	72.044	5.402	44.8	69.3	72.2	75.4	90

2.7.2. Exploratory data analysis

After analyzing Figure 3(a), the dataset is composed of a total of 615 data, of which 377 correspond to suspected hepatitis C patients and 238 to healthy patients. This distribution provides relevant information on the proportion of healthy in the sample. Similarly in Figure 3(b) the data indicate a gender distribution in the sample in which males represent 61.30% of the total, while females constitute 38.70%.



(a) (b)

Figure 3. Percentage ratio between (a) healthy and suspicious patients and (b) patients by gender

Meanwhile, analyzing in Figure 4(a) most of the patients' show CHE levels concentrated in the range of 7 to 10, suggesting a significant prevalence of values in that range. Figure 4(b) CHOL, most patients show levels in the range 4 to 7. Figure 4(c) ALP, most patients' show values ranging from 40 to 90, while for Figure 4(d) ALT is mainly in the range of 10 to 25.

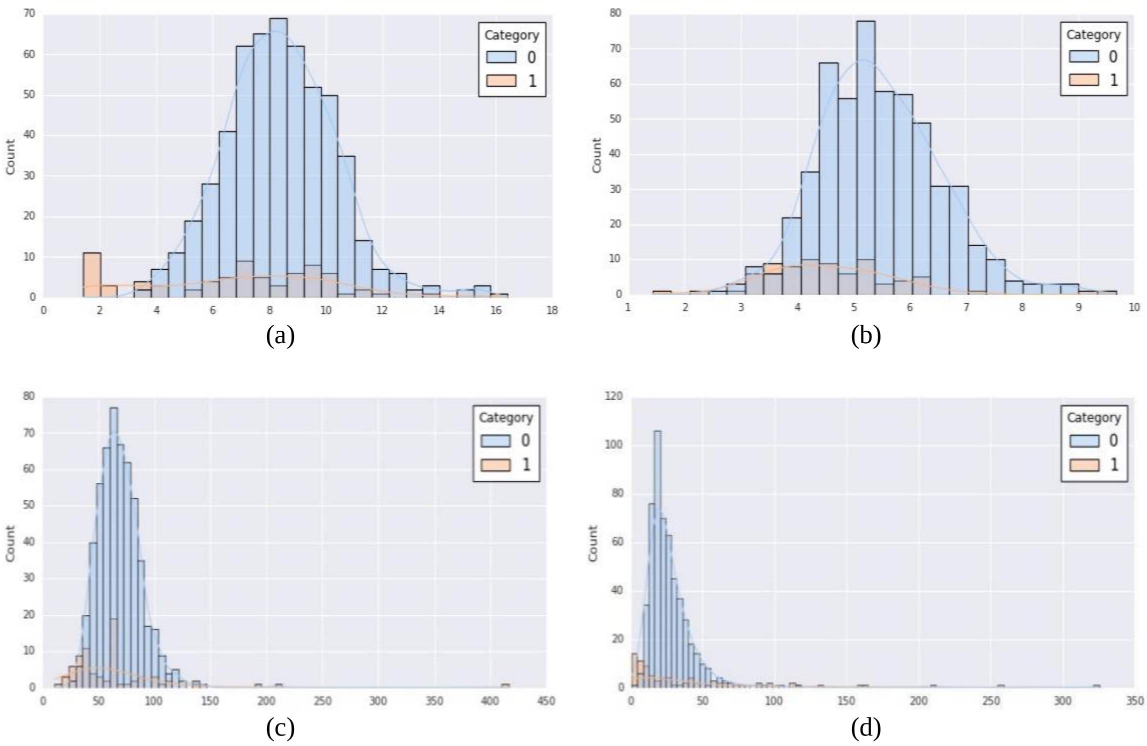
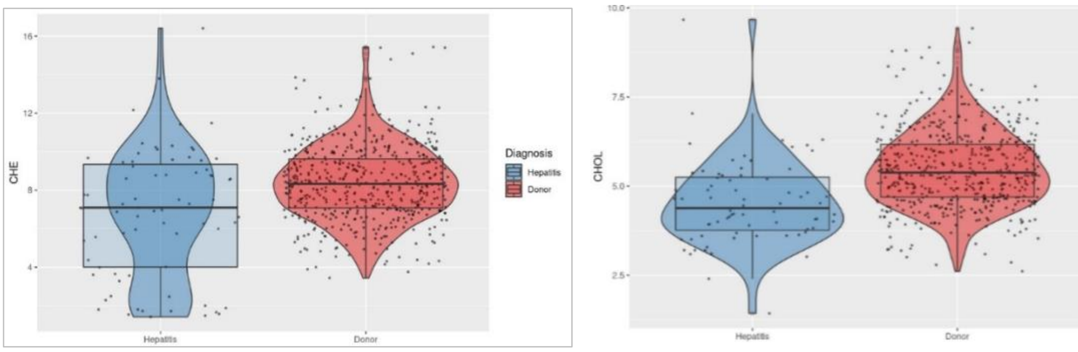


Figure 4. Levels of substances present in the human body for (a) CHE, (b) total CHOL, (c) ALP, and (d) ALT

In Figure 5, an analysis of certain variables is performed to identify possible relationships that may be linked to the likelihood of contracting HCV. When analyzing Figure 5(a), most HCV patients have CHE levels in the range of 4 to 9, in contrast to those without HCV disease, whose CHE levels generally range between 7 and 9. Similarly, in Figure 5(b), it is noted that most HCV patients have CHOL levels in the range of 3.5 to 5.



(a) (b)

Figure 5. Relationship of the variable (a) CHE levels and (b) HCV patients

In the correlation of variables, the correlation between ALB and PROT is 0.5, suggesting a moderate positive correlation between these two blood components. In addition, ALB and CHE correlate 0.4, indicating a moderate positive relationship. Likewise, ALP and GGT share a correlation of 0.4, and AST and GGT correlate 0.5. Finally, CHE and CHOL correlate 0.4. Overall, these correlations suggest that there are significant associations between these variables in the dataset. However, values of -0.2 for the correlation between AST and CHOL, as well as AST and CHE, indicate a moderate negative correlation. This means that when AST levels increase, CHOL and CHE levels tend to decrease. Also, a value of -0.1 for the correlation between the variables "sex" and "ALB" suggests a weak negative correlation. This means that, in the dataset, there is a minimal relationship between the sex of individuals and their blood ALB levels.

2.7.3. Data processing

Before starting the model training process, it is necessary to perform data processing and debugging to optimize the performance of the algorithms. In this data processing stage, one of the crucial steps is to divide the dataset into training and test sets. In this case, the focus was on predicting whether a patient has HCV or not. To do this, the category variable to be predicted was designated as "y" and the rest of the data, which constitute the medical characteristics, as "X" or the input data. This division allows training the models with one part of the data and then assessing their predictive ability using the other part. After data preparation, several ML models were selected, including LR, RF, KNN, DT, catBoost classifier (CATBC), and GB. The choice of these models was made to carry out a comparative evaluation of their performance in HCV prediction.

3. RESULTS AND DISCUSSION

Following the data preparation and pre-processing process for HCV detection, several ML models were trained to determine the most efficient in terms of accuracy, precision, sensitivity, and F1 score. The models tested were LR, RF, KNN, DT, CATBC, and GB. The results of this training process are specified in Table 3, which provides a comprehensive overview of the performance of each model about the metrics.

After completion of the training stage, the LR, RF, KNN, DT, CATBC, and GB algorithms achieved accuracy rates of 89%, 93%, 85%, 93%, 93%, 92%, and 94%, in that order. Furthermore, according to Table 3, it is observed that the GB model excels in terms of accuracy, sensitivity, F1 score, and mean accuracy, reaching values of 94%, 97%, 85%, and 90%, respectively. This places it as the most effective predictor for HCV detection. The second-best performance corresponds to the RF and DT models, with values of 93% in accuracy, 93% in sensitivity, 93% in F1 score, and 92% in mean accuracy. In third place is the slightly lower-performing CATBC model, with 92% accuracy, 93% sensitivity, 92% F1 score, and 91% average accuracy. The LR model is in fourth place, with an accuracy of 89%, sensitivity of 88%, F1 score of 89%, and an average accuracy of 87%. Finally, the KNN model has the least favorable indicators, with 85% precision, 88% sensitivity, 85% F1 score, and 82% average accuracy.

HCV is a virus that affects the liver and is transmitted mainly through contact with the blood of an infected person. Around 58 million people are chronically infected with HCV, with approximately 1.5 million new infections each year. In 2019, 290,000 people lost their lives to the disease. It is therefore essential to conduct a study to evaluate and contrast several ML models to determine which one provides the highest levels of accuracy in predicting HCV. The LR, RF, KNN, DT, CATBC, and GB models were trained. The results indicated that the GB algorithm achieved the highest score, reaching an accuracy rate of 94%. This differs from Ma *et al.* [31] where they found XGBoost to be the best predictor with an accuracy of 91.56%, which is lower than the accuracy achieved in this study of 94% with GB. On the other hand, Islam *et al.* [32] ranked ANN as the most accurate with an accuracy of 95.50%. In comparison, this study achieved a lower accuracy, which is probably due to the volume of the dataset used, as well as the techniques used. Contrary to Hafeez *et al.* [33] which used a different dataset than this study and achieved a lower accuracy with SVM of 91.84%. Similar to Rouhani and Haghighi [34] where the SVM model together with ANN achieved an accuracy of 97%. The optimization strategies employed often have an impact on these results. On the other hand, several researchers [36], [39], positioned RF as the most efficient predictor for HCV with an accuracy of 99.46% and 98.6% respectively, surpassing that obtained in this work. ML models can make a significant contribution to HCV detection, but their effectiveness is highly dependent on the quality of the datasets used and the optimization techniques and strategies applied to the models.

Table 3. Model training results

	Precision (%)	Recall (%)	F1-score (%)	Support
		LR		
0	89	98	93	99
1	86	50	63	24
accuracy			89	123
macro avg	87	74	78	123
weighted avg	88	89	87	123
		RF		
0	92	99	96	99
1	94	67	78	24
accuracy			93	123
macro avg	93	83	87	123
weighted avg	93	93	92	123
		KNN		
0	85	100	92	99
1	100	25	40	24
accuracy			85	123
macro avg	92	62	66	123
weighted avg	88	85	82	123
		DT		
0	92	99	96	99
1	94	67	78	24
accuracy			93	123
macro avg	93	83	87	123
weighted avg	93	93	92	123
		CATBC		
0	91	100	95	99
1	100	58	74	24
accuracy			92	123
macro avg	95	79	84	123
weighted avg	93	92	91	123
		GB		
0	93	100	97	99
1	100	71	83	24
accuracy			94	123
macro avg	97	85	90	123
weighted avg	95	94	94	123

4. CONCLUSION

In this study, the potential of ML models to predict the presence of HCV was explored. An evaluation of the accuracy of these models in detecting HCV was carried out. After presenting the results obtained by training the LR, RF, KNN, DT, CATBC, and GB models on the task of HCV prediction, the following conclusions have been reached. The GB model showed outstanding performance, obtaining the strongest metrics in terms of precision, accuracy, and sensitivity in HCV detection. This model could be essential for the early detection of HCV. The second-best performance was attributed to the RF and DT models, which achieved 93% accuracy. In third place was the CATBC model with an accuracy of 92%. In fourth position was the LR model, which achieved an accuracy of 89%. Finally, the KNN model exhibited the least favorable results, with an accuracy of 85%. To increase the efficiency and robustness of ML models in future research, it is recommended to consider the implementation of a variety of optimization techniques, as well as the incorporation of additional and more diversified datasets. These strategies can significantly contribute to improving the performance of the models in the HCV prediction task and may be a key aspect in future advances in this field of study. In addition, it is recommended to explore the performance of other ML algorithms, such as SVM, neural networks, or model ensembles. Also, it would be interesting to develop future work on how to implement these models in clinical practice and evaluate their impact on patient care. Finally, the models have proven to be a reliable tool in HCV identification, suggesting their potential utility in clinical trials. However, it is crucial to keep in mind that their efficacy is closely related to the quality and quantity of the data used and the optimization strategies implemented.

REFERENCES

[1] WHO, "WHO publishes updated guidance on hepatitis C infection – with new recommendations on treatment of adolescents and children, simplified service delivery and diagnostics," *World Health Organization*, 2022. Accessed: Oct. 08, 2023. [Online]. Available: https://www.who.int/news/item/24-06-2022-WHO-publishes-updated-guidance-on-hepatitis-C-infection

[2] WHO, "Hepatitis C," *World Health Organization*, 2023. Accessed: Oct. 08, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/hepatitis-c

[3] G. Nourse, A. Farrugia, S. Fraser, D. Moore, and C. Treloar, "Optimism and eternal vigilance: gathering disease, responsible subjects and the hope of elimination in the new hepatitis C treatment era," *International Journal of Drug Policy*, vol. 119, 2023, doi: 10.1016/j.drugpo.2023.104142.

[4] I. O. Adedotun *et al.*, "Molecular docking, ADMET analysis, and bioactivity studies of phytochemicals from Phyllanthus niruri as potential inhibitors of hepatitis C virus NSB5 polymerase," *Journal of the Indian Chemical Society*, vol. 99, no. 2, 2022, doi: 10.1016/j.jics.2021.100321.

[5] R. Ingle, A. K. Chaya, S. Chavan, S. Taklikar, and S. Baveja, "A study of seroprevalence and the associated risk factors of hepatitis C at a tertiary care hospital in Mumbai," *Clinical Epidemiology and Global Health*, vol. 23, 2023, doi: 10.1016/j.cegh.2023.101356.

[6] P. K. Parikh *et al.*, "Developments in small molecule antiviral drugs against hepatitis B and C viruses: FDA approved therapies and new drugs in clinical trials," *Arabian Journal of Chemistry*, vol. 16, no. 8, 2023, doi: 10.1016/j.arabjc.2023.105013.

[7] S. M. Walters *et al.*, "How the rural risk environment underpins hepatitis C risk: qualitative findings from rural southern Illinois, United States," *International Journal of Drug Policy*, vol. 112, 2023, doi: 10.1016/j.drugpo.2022.103930.

[8] A. H. Saputro *et al.*, "Alpha-mangostin, piperine, and beta-sitosterol as hepatitis C antivirus (HCV): in silico and in vitro studies," *Heliyon*, vol. 9, no. 9, 2023, doi: 10.1016/j.heliyon.2023.e20141.

[9] V. Jaiswal *et al.*, "Cardioprotective effect of antiviral therapy among hepatitis C infected patients: a meta-analysis," *IJC Heart and Vasculature*, vol. 49, 2023, doi: 10.1016/j.ijcha.2023.101270.

[10] B. B. Warssamo and D. B. Belay, "Knowledge, attitude and practice of hepatitis C virus among waste handlers in Sidama, Ethiopia," *Scientific African*, vol. 21, 2023, doi: 10.1016/j.sciaf.2023.e01764.

[11] J. Demant, L. K. -Dehli, J. V. D. Veen, A. Øvrehus, J. V. Lazarus, and N. Weis, "Peer-delivered point-of-care testing and linkage to treatment for hepatitis C virus infection among marginalized populations through a mobile clinic in Copenhagen, Denmark," *International Journal of Drug Policy*, vol. 121, 2023, doi: 10.1016/j.drugpo.2023.104185.

[12] S. Dröse, A. L. H. Øvrehus, D. K. Holm, B. T. Røge, and P. B. Christensen, "Hepatitis C screening and linkage to care with a mobile clinic in Southern Denmark," *International Journal of Drug Policy*, vol. 121, 2023, doi: 10.1016/j.drugpo.2023.104180.

[13] S. Mahmud *et al.*, "Efficiency and safety of sofosbuvir in Bangladeshi children with chronic hepatitis C virus infection," *iLIVER*, vol. 2, no. 3, pp. 146–150, 2023, doi: 10.1016/j.iliver.2023.06.002.

[14] Y. A. -Meléndez *et al.*, "Hepatitis C virus care cascade among people who inject drugs in Puerto Rico: minimal HCV treatment and substantial barriers to HCV care," *Drug and Alcohol Dependence Reports*, vol. 8, p. 100178, 2023, doi: 10.1016/j.dadr.2023.100178.

[15] A. K. Martin *et al.*, "Peer recovery coaching for comprehensive HIV, hepatitis C, and opioid use disorder management: the chorus pilot study," *Drug and Alcohol Dependence Reports*, vol. 7, 2023, doi: 10.1016/j.dadr.2023.100156.

[16] WHO, "WHO launches 'one life, one liver' campaign on world hepatitis day," *World Health Organization*, 2023. Accessed: Oct. 14, 2023. [Online]. Available: https://www.who.int/news/item/28-07-2023-who-launches--one-life--one-liver--campaign-on-world-hepatitis-day

[17] O. A. R. AboZaid *et al.*, "Sofosbuvir plus ribavirin combination regimen boost liver functions and antioxidant profile in hepatitis C virus patients," *Microbial Pathogenesis*, vol. 150, 2021, doi: 10.1016/j.micpath.2021.104740.

[18] D. K. V. Santen *et al.*, "Treatment as prevention effect of direct-acting antivirals on primary hepatitis C virus incidence: findings from a multinational cohort between 2010 and 2019," *eClinicalMedicine*, vol. 56, 2023, doi: 10.1016/j.eclinm.2022.101810.

[19] C. Lanièce Delaunay *et al.*, "Public health interventions, priority populations, and the impact of COVID-19 disruptions on hepatitis C elimination among people who have injected drugs in Montreal (Canada): a modeling study," *International Journal of Drug Policy*, vol. 116, 2023, doi: 10.1016/j.drugpo.2023.104026.

[20] N. Zaman *et al.*, "Hepatitis C virus infection in garbage pickers of different districts of Khyber Pakhtunkhwa, Pakistan," *Dialogues in Health*, vol. 1, 2022, doi: 10.1016/j.dialog.2022.100073.

[21] W. W. L. Wong *et al.*, "Time costs and out-of-pocket costs in patients with chronic hepatitis C in a publicly funded health system," *Value in Health*, vol. 25, no. 2, pp. 247–256, 2022, doi: 10.1016/j.jval.2021.08.006.

[22] A. C. -Villalvir, J. M. Wilkerson, C. Markham, L. Rodriguez, and V. Schick, "A qualitative investigation of the barriers and facilitators to hepatitis C virus (HCV) screening among individuals experiencing homelessness in Houston, Texas," *Dialogues in Health*, vol. 1, 2022, doi: 10.1016/j.dialog.2022.100058.

[23] K. Swetha, A. Kiran, K. Pavanam, E. N. V. Kumari, T. Naresh, and M. J. Baba, "Inflammation of liver and hepatitis disease prediction using machine learning techniques," in *Proceedings of the 7th International Conference on Intelligent Computing and Control Systems, ICICCS 2023*, 2023, pp. 218–223, doi: 10.1109/ICICCS56967.2023.10142912.

[24] E. Y. -N. Kang, D. R. Chen, and Y. Y. Chen, "Associations between literacy and attitudes toward artificial intelligence–assisted medical consultations: the mediating role of perceived distrust and efficiency of artificial intelligence," *Computers in Human Behavior*, vol. 139, 2023, doi: 10.1016/j.chb.2022.107529.

[25] A. Alotaibi *et al.*, "Explainable ensemble-based machine learning models for detecting the presence of cirrhosis in hepatitis C patients," *Computation*, vol. 11, no. 6, 2023, doi: 10.3390/computation11060104.

[26] J. Price, T. Yamazaki, K. Fujihara, and H. Sone, "XGBoost: interpretable machine learning approach in medicine," *WSCE 2022 - 2022 5th World Symposium on Communication Engineering*, pp. 109–113, 2022, doi: 10.1109/WSCE56210.2022.9916029.

[27] M. Aamir, S. Bazai, U. A. Bhatti, Z. A. Dayo, J. Liu, and K. Zhang, "Applications of machine learning in medicine: current trends and prospects," *2023 Global Conference on Wireless and Optical Technologies (GCWOT)*, Malaga, Spain, 2023, pp. 1-4, doi: 10.1109/GCWOT57803.2023.10064665.

[28] T. Oguguo, G. Zamzmi, S. Rajaraman, F. Yang, Z. Xue, and S. Antani, "A comparative study of fairness in medical machine learning," *Proceedings - International Symposium on Biomedical Imaging*, pp. 1–5, 2023, doi: 10.1109/ISBI53787.2023.10230368.

[29] Z. S. Hossein Abad, A. Kline, and J. Lee, "Evaluation of machine learning-based patient outcome prediction using patient-specific difficulty and discrimination indices," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2020, pp. 5446–5449, doi: 10.1109/EMBC44109.2020.9176622.

[30] R. Kumar, P. Thakur, and S. Chauhan, "Special disease prediction system using machine learning," in *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing, COM-IT-CON 2022*, 2022, pp. 42–45, doi: 10.1109/COM-IT-CON54601.2022.9850843.

[31] L. Ma, Y. Yang, X. Ge, Y. Wan, and X. Sang, "Prediction of disease progression of chronic hepatitis C based on XGBoost algorithm," in *2020 International Conference on Robots and Intelligent Systems, ICRIS 2020*, 2020, pp. 598–601, doi: 10.1109/ICRIS52159.2020.00151.

[32] S. Islam, A. U. Rehman, S. Javaid, T. M. Ali, and A. Nawaz, "An integrated machine learning framework for classification of cirrhosis, fibrosis, and hepatitis," *2022 Third International Conference on Latest trends in Electrical Engineering and Computing Technologies (INTELLECT)*, Karachi, Pakistan, 2022, pp. 1-6, doi: 10.1109/INTELLECT55495.2022.9969404.

[33] M. A. Hafeez, A. Imran, M. I. Khan, A. H. Khan, A. Nawaz, and S. Ahmed, "Diagnosis of liver disease induced by hepatitis virus using machine learning methods," in *8th International Conference on Information Technology Trends: Industry 4.0: Technology Trends and Solutions, ITT 2022*, 2022, pp. 154–159, doi: 10.1109/ITT56123.2022.9863944.

[34] M. Rouhani and M. M. Haghighi, "The diagnosis of hepatitis diseases by support vector machines and artificial neural networks," in *2009 International Association of Computer Science and Information Technology - Spring Conference, IACSIT-SC 2009*, 2009, pp. 456–458, doi: 10.1109/IACSIT-SC.2009.25.

[35] S. O. Olatunji *et al.*, "Preemptive diagnosis of Hepatitis c using machine learning techniques: a retrospective study in Saudi Arabia," *2023 International Conference on Smart Computing and Application (ICSCA)*, Hail, Saudi Arabia, 2023, pp. 1-6, doi: 10.1109/ICSCA57840.2023.10087834.

[36] T. A. N. Saputra, K. I. Arizona, M. R. Andrian, F. I. Kurniadi, and B. Juarto, "Random forest in detecting hepatitis C," in *2022 9th International Conference on Information Technology, Computer and Electrical Engineering, ICITACEE 2022*, 2022, pp. 299–302, doi: 10.1109/ICITACEE55701.2022.9924074.

[37] K. R. Singh, R. Gupta, R. K. Kadian, and R. Singh, "An optimized XGboost approach for predicting progression of hepatitis C using hyperparameter tuning and feature interaction constraint," *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)*, Ravet, India, 2022, pp. 1-8, doi: 10.1109/ASIANCON55314.2022.9909086.

[38] M. Shahzadi, F. Bukhari, and N. Shafi, "Intelligent predictive model for hepatitis C," in *3rd IEEE International Conference on Artificial Intelligence, ICAI 2023*, 2023, pp. 7–12, doi: 10.1109/ICAI58407.2023.10136685.

[39] T. I. Trishna, S. U. Emon, R. R. Ema, G. I. H. Sajal, S. Kundu, and T. Islam, "Detection of hepatitis (A, B, C, and E) viruses based on random forest, k-nearest and naïve Bayes classifier," *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Kanpur, India, 2019, pp. 1-7, doi: 10.1109/ICCCNT45670.2019.8944455.

[40] L. Wang, T. Wang, and X. Hu, "Logistic regression region weighting for weakly supervised object localization," *IEEE Access*, vol. 7, pp. 118411–118421, 2019, doi: 10.1109/ACCESS.2019.2935011.

[41] A. B. Amjoud and M. Amrouch, "Transfer learning for automatic image orientation detection using deep learning and logistic regression," *IEEE Access*, vol. 10, pp. 128543–128553, 2022, doi: 10.1109/ACCESS.2022.3225455.

[42] J. C. Nwadiuto, S. Yoshino, H. Okuda, and T. Suzuki, "Variable selection and modeling of drivers' decision in overtaking behavior based on logistic regression model with gazing information," *IEEE Access*, vol. 9, pp. 127672–127684, 2021, doi: 10.1109/ACCESS.2021.3111753.

[43] D. Lei, J. Tang, Z. Li, and Y. Wu, "Using low-rank approximations to speed up kernel logistic regression algorithm," *IEEE Access*, vol. 7, pp. 84242–84252, 2019, doi: 10.1109/ACCESS.2019.2924542.

[44] B. Wang and J. Zhang, "Logistic regression analysis for lncrna-disease association prediction based on random forest and clinical stage data," *IEEE Access*, vol. 8, pp. 35004–35017, 2020, doi: 10.1109/ACCESS.2020.2974624.

[45] L. Dong *et al.*, "Very high resolution remote sensing imagery classification using a fusion of random forest and deep learning technique-subtropical area for example," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 113–128, 2020, doi: 10.1109/JSTARS.2019.2953234.

[46] P. Josso, A. Hall, C. Williams, T. Le Bas, P. Lusty, and B. Murton, "Application of random-forest machine learning algorithm for mineral predictive mapping of Fe-Mn crusts in the world ocean," *Ore Geology Reviews*, vol. 162, 2023, doi: 10.1016/j.oregeorev.2023.105671.

[47] S. Latifi, D. Jannach, and A. Ferraro, "Sequential recommendation: a study on transformers, nearest neighbors and sampled metrics," *Information Sciences*, vol. 609, pp. 660–678, 2022, doi: 10.1016/j.ins.2022.07.079.

[48] Z. Xu, J. Cao, G. Zhang, X. Chen, and Y. Wu, "Active learning accelerated Monte-Carlo simulation based on the modified K-nearest neighbors algorithm and its application to reliability estimations," *Defence Technology*, vol. 28, pp. 306–313, 2023, doi: 10.1016/j.dt.2022.09.012.

[49] A. X. Wang, S. S. Chukova, and B. P. Nguyen, "Ensemble k-nearest neighbors based on centroid displacement," *Information Sciences*, vol. 629, pp. 313–323, 2023, doi: 10.1016/j.ins.2023.02.004.

[50] S. Suyanto, P. E. Yunanto, T. Wahyuningrum, and S. Khomsah, "A multi-voter multi-commission nearest neighbor classifier," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, pp. 6292–6302, 2022, doi: 10.1016/j.jksuci.2022.01.018.

[51] O. I. -Villanueva, K. E. -Linares, R. O. F. Castañeda, and M. C. -Carbonell, "Application of machine learning models for early detection and accurate classification of type 2 diabetes," *Diagnostics*, vol. 13, no. 14, Jul. 2023, doi: 10.3390/diagnostics13142383.

[52] N. Kumar, M. M. Tripathi, S. Gupta, M. A. Alotaibi, H. Malik, and A. Afthanorhan, "Study of potential impact of wind energy on electricity price using regression techniques," *Sustainability*, vol. 15, no. 19, 2023, doi: 10.3390/su151914448.

[53] F. Masood, W. U. Khan, S. U. Jan, and J. Ahmad, "AI-enabled traffic control prioritization in software-defined iot networks for smart agriculture," *Sensors*, vol. 23, no. 19, 2023, doi: 10.3390/s23198218.

[54] J. C. Zaveri, S. R. Dhanushkodi, C. R. Kumar, J. Taler, M. Majdak, and B. Węglowski, "Predicting the performance of pem fuel cells by determining dehydration or flooding in the cell using machine learning models," *Energies*, vol. 16, no. 19, Oct. 2023, doi: 10.3390/en16196968.

[55] I. D. Mienye and Y. Sun, "A survey of ensemble learning: concepts, algorithms, applications, and prospects," *IEEE Access*, vol. 10, pp. 99129–99149, 2022, doi: 10.1109/ACCESS.2022.3207287.

[56] P. Duan *et al.*, "High-resolution planetscope imagery and machine learning for estimating suspended particulate matter in the Ebinur Lake, Xinjiang, China," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 1019–1032, 2023, doi: 10.1109/JSTARS.2022.3233113.
[57] S. K. Hashemi, S. L. Mirtaheri, and S. Greco, "Fraud detection in banking data by machine learning techniques," *IEEE Access*, vol. 11, pp. 3034–3043, 2023, doi: 10.1109/ACCESS.2022.3232287.
[58] J. Li, H. Zhang, and Z. Wei, "The weighted word2vec paragraph vectors for anomaly detection over http traffic," *IEEE Access*, vol. 8, pp. 141787–141798, 2020, doi: 10.1109/ACCESS.2020.3013849.
[59] M. T. Vo, T. Nguyen, and T. Le, "Robust head pose estimation using extreme gradient boosting machine on stacked autoencoders neural network," *IEEE Access*, vol. 8, pp. 3687–3694, 2020, doi: 10.1109/ACCESS.2019.2962974.
[60] F. Alzamzami, M. Hoda, and A. E. Saddik, "Light gradient boosting machine for general sentiment classification on short texts: a comparative evaluation," *IEEE Access*, vol. 8, pp. 101840–101858, 2020, doi: 10.1109/ACCESS.2020.2997330.
[61] Y. Shao and C. Wang, "HIBoosting: a recommender system based on a gradient boosting machine," *IEEE Access*, vol. 7, pp. 171013–171022, 2019, doi: 10.1109/ACCESS.2019.2956342.

BIOGRAPHIES OF AUTHORS

Orlando Iparraguirre-Villanueva is systems engineer with a master's degree in information technology management, Ph.D. in Systems Engineering from the Universidad Nacional Federico Villarreal, Peru. ITIL® Certified, Specialization in Business Continuity Management (SBCM), SCRUM Certified. National and international speaker/panelist (Panama, Colombia, Ecuador, Venezuela, Mexico). Extensive experience in undergraduate and postgraduate teaching in different universities in the country. Thesis advisor and jury in different universities. Professional experience in management positions in the field of Information Technology. Research professor with publications in Scopus and WoS-indexed journals (Q1, Q2, Q3, and Q4) of high impact. Topics of interest: open-source software, IoT, augmented reality, machine learning, AI, CNN, text mining, virtual environments, scientific research methodology, and thesis. He can be contacted at email: oiparraguirre@ieee.org.

Rosalynn Ornella Flores-Castañeda systems engineer with master's degrees in administration and international Relations (2011) and in Information Technology Management (2016). Ph.D. in Educational Administration (2014) and in Systems Engineering (2021). Former Director of Academic Records at the Universidad César Vallejo, teacher, and thesis advisor. Member of the College of Engineers of Peru. Author of educational and engineering publications. She can be contacted at email: rfloresc@ucv.edu.pe.

Henry Chero-Valdivieso degree in Mathematics Education and a master's degree in education. Ph.D. student in Communication and Education in digital environments. Specialist in ICT and postgraduate studies in Systems Engineering. Extensive teaching experience in mathematics and technologies, as well as in ICT projects for the Ministry of Education in Peru. Specialist in e-learning and consultant in educational information technologies. He can be contacted at email: hacheroc@ucvvirtual.edu.pe.

Fernando Sierra-Liñan has a bachelor's degree in education, specializing in Science and Technology at USIL, a master's degree in Edumatics and University Teaching at UTP, a bachelor's degree in systems Engineering and Computer Science at UTP, with a technical specialty in Computer Science and Computer Science. He is currently working as a researcher and thesis advisor in the faculty of Computer Engineering and Systems at the Universidad Privada del Norte, Lima, Peru. He has 20 years of teaching experience. Additionally, he teaches the research methodology course at the Cesar Vallejo University, Lima, Peru. His areas of interest are programming, database, and data analysis. He can be contacted at email: fsierra@ucvvirtual.edu.pe.