Machine Learning Algorithm for Predicting Hepatocellular Carcinoma in HCV Patients

Kamel Maaloul , Brahim Lejdel * , Eliseo Clementini

Posted Date: 9 February 2026

doi: 10.20944/preprints202602.0659.v1

Keywords: hepatocellular carcinoma (HCC); machine learning; hepatitis C virus (HCV); gradient boosting classifier

 doi:10.20944/preprints202602.0659.v1

Article

Machine Learning Algorithm for Predicting Hepatocellular Carcinoma in HCV Patients

Kamel Maaloul [1], Brahim Lejdel [1] and Eliseo Clementini [2]

[1] Computer Science Department, Algeria

[2] Department of Industrial and Information Engineering and Economics, University of El-Oued, University of L'Aquila, L'Aquila, Italy

* Correspondence: lejdel.brahim@gmail.com

Abstract

Background/Objectives: Hepatocellular carcinoma (HCC) is a leading cause of cancer-related mortality worldwide and is frequently associated with chronic hepatitis C virus (HCV) infection. Early prediction of HCC in HCV patients remains challenging due to complex clinical patterns. This study aims to develop and evaluate machine learning models for the early prediction of hepatocellular carcinoma in patients with HCV. **Methods:** Clinical and laboratory data from HCV patients were analyzed using a machine learning–based framework. The dataset was preprocessed, and relevant features were selected prior to model development. Six supervised machine learning algorithms—CatBoost, XGBoost, LightGBM, Gaussian Naive Bayes, Extra Trees, and Random Forest—were implemented. Hyperparameter optimization was performed using the Optuna framework. Model performance was assessed using standard evaluation metrics, including accuracy, precision, recall, and F1-score. **Results**: The experimental results demonstrate that machine learning techniques can effectively identify patterns associated with the progression to hepatocellular carcinoma in HCV patients. Among the evaluated models, ensemble-based algorithms achieved the highest predictive performance, outperforming baseline approaches across multiple evaluation metrics. **Conclusions:** The findings confirm that machine learning models can serve as valuable decision-support tools for the early detection of hepatocellular carcinoma in patients with HCV. Integrating such models into clinical workflows may enhance early diagnosis and improve patient outcomes. Future work will focus on expanding the dataset and validating the models in real-world clinical settings.

Keywords: hepatocellular carcinoma (HCC); machine learning; hepatitis C virus (HCV); gradient boosting classifier

1. Introduction

The most prevalent primary liver cancer and one of the top four causes of cancer-related mortality globally is hepatocellular carcinoma (HCC). HCC is the leading cause of cancer-related fatalities in the United States, and by 2030, it may overtake all other causes. Antiviral therapy and the hepatitis B vaccine have been utilized extensively in recent years, and there are several different ways to treat HCC, including as surgery, ablation, transcatheter arterial chemoembolization (TACE), chemotherapy targeted immunotherapy, and others. HCC has a bad prognosis overall, and its signs are difficult to identify. Early detection, precise prediction, customized treatment, and follow-up are all necessary for HCC [1].

Chronic liver disease (CLD) is a significant global health concern, responsible for approximately 2 million deaths each year. The primary causes of liver damage include chronic alcohol consumption, immune system disorders, and liver infections caused by hepatitis B and C viruses. Such damage triggers a wound-healing response in the liver, which may progress to cirrhosis, hepatic fibrosis (HF), and, ultimately, liver failure or hepatic cancer. Among these causes, hepatitis C virus (HCV) infection

 doi:10.20944/preprints202602.0659.v1

stands out as the leading trigger of chronic liver inflammation and a key risk factor for hepatocellular carcinoma (HCC) [2]. HCC, a malignant tumor of the liver, holds the distinction of being the fifth most common cancer globally and the third leading cause of cancer-related deaths worldwide. Addressing this issue, the World Health Organization (WHO) launched its first global health strategy on viral hepatitis in June 2016. This initiative aims to reduce the incidence and mortality rates of chronic viral hepatitis (CVH) by 90% and 65%, respectively, by 2030 [3].

Artificial intelligence (AI) represents an emerging field that explores and develops theories, methodologies, technologies, and application systems aimed at simulating, extending, and enhancing human intelligence. At the core of AI lies machine learning (ML), a discipline that enables computers to emulate or replicate human learning behaviors, allowing them to acquire new knowledge or skills and refine existing information to enhance their performance. Over the past decade, ML has been increasingly adopted in medical research, leading to notable advancements across various domains [4]. This progress is particularly evident in cancer-related inquiries, encompassing areas like lung cancer, breast cancer, and hepatocellular carcinoma (HCC). Studies applying ML to HCC address multiple aspects including diagnosis, treatment, prognosis, and algorithm model development. The integration of ML in HCC research not only sheds light on the interplay between AI and HCC but also contributes significantly to its prevention and treatment strategies. This review emphasizes the role of ML in advancing HCC-related diagnosis, therapeutic approaches, and prognostic evaluations [5].

The sole method for verifying the clinical measurement of HCC is a liver biopsy. However, it has several disadvantages and significant problems (e.g., pain in the muscles, hemorrhage, contamination, harm to nearby organs, inaccurate diagnosis, and variability between and among observers) [6]. A less invasive and more accurate way to assess the degree and course of HCC would be very helpful in a clinical setting [7]. Basic scoring methods have so far been developed using different combinations of common clinical features, ignoring population heterogeneity and viral eradication. By identifying risk variables in patients with other chronic illnesses, machine learning algorithms can aid in the prediction of HCC [8].

Machine learning is a complete tool for model creation that has surfaced in recent years. It allows for automatic selection of predicting factors and uses maximum data to reduce bias [9]. Novel clinical, radiological, and pathological features-based prediction models utilizing machine learning algorithms can be created to ascertain the risk levels of HCC in patients with HCV [10]. By incorporating these models into computer-based management systems, it may be possible to enhance the clinical assessment and risk classification of HCC in patients infected with the hepatitis C virus [11]. The Gradient Boosting-based strategy is one machine learning technique [12].

The remainder of this paper is structured as follows. Chapter 2 reviews previous studies that investigated hepatocellular carcinoma prediction in HCV patients using different machine learning techniques and analytical approaches. Chapter 3 presents the dataset and outlines the methodology adopted to develop the proposed artificial intelligence–based model. Chapter 4 reports the experimental evaluation and discusses the obtained results, followed by conclusions and directions for future research. Finally, the paper concludes with a list of references.

2. Related Work

The world today depends on data, so artificial intelligence plays a major role in smart cities. Researchers have successfully applied several algorithms and models to predict Predicting Hepatocellular Carcinoma in HCV Patients conditions using various metrological characteristics, features, and data generated from different sources.

In order to determine whether cirrhosis is present in hepatitis C patients, Alotaibi, A., et al. [13] employed ensemble-based machine learning models. The data, which consisted of 28 characteristics and included 2038 patients from Egypt, was used to train four machine learning models: Gradient Boosting Machine, RF, ExtraT, and XGBoost. The ExtraT model outperformed the other two models,

 doi:10.20944/preprints202602.0659.v1

according to the results, with 96.92% accuracy, 99.81% precision, and 94.00% recall using 16 of the 28 features.

The XGBoost algorithm was employed by Ma, L. et al. [14] to predict the presence of hepatitis C in clinical datasets from chronic hepatitis C patients and blood donors. Following model evaluation, the experimental findings show that the XGBoost model is reliable and outperforms the Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), and Adaboost algorithms, with an accuracy of 91.56%. Similarly, [20] studied the classification, diagnosis, and prediction of hepatitis C using four (04) machine learning techniques: KNN, SVM, Naive Bayes, and DT. With an accuracy of 93.44%, the DT approach outperformed the other models in terms of the particular classification objectives, according to the study of 615 data.

In Oladimeji et al. [15], the authors used the SMOTE technique (Synthetic Minority Oversampling Technique) to address the imbalance issue in the dataset that was taken from the UCI-ML Repository. According to their findings, RF outperformed all other evaluation parameters. [16] and [17] used the XGBoost model to predict hepatitis C illness with less than 90% accuracy.

3. Proposed Architecture

In this paper, the proposed model for Predicting Hepatocellular Carcinoma in HCV Patients prediction using several machine learning algorithms mainly aims to collect data related to parameters.

The chart below is the proposed model for forecasting the Predicting Hepatocellular Carcinoma in HCV Patients using several machine learning algorithms.

Model creation, model optimization, model evaluation, data collecting, data pre-processing, data splitting, and deployment of the optimal model are the six steps that make up the proposed framework. The suggested paradigm for forecasting HCV illness is depicted in Figure 1.

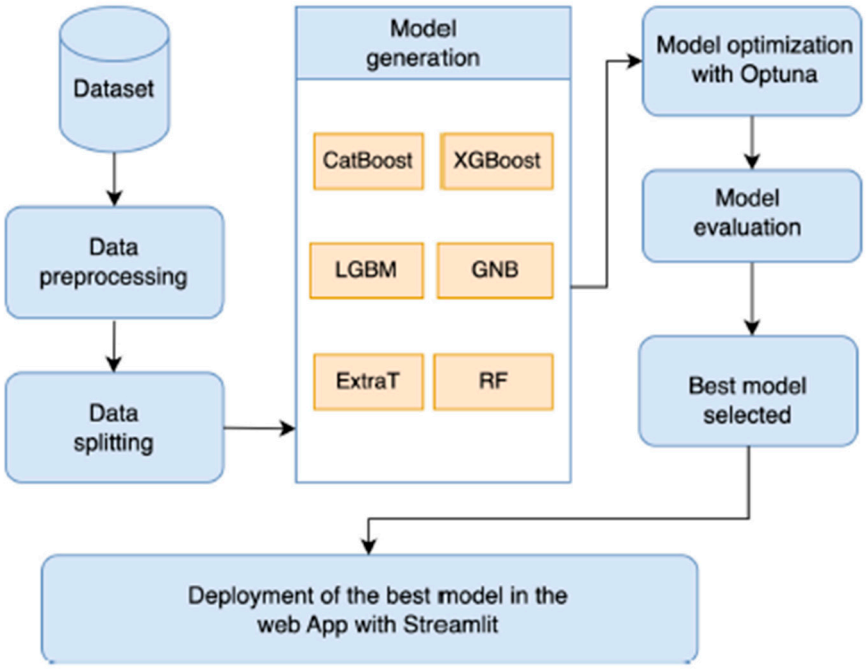

Figure 1. A suggested framework for detecting hepatitis C. [23].

The proposed framework follows a sequence of well-defined stages. First, the dataset is collected and prepared for use in subsequent phases. Next, data pre-processing techniques are applied to clean

 doi:10.20944/preprints202602.0659.v1

and transform the raw data prior to model development. The processed dataset is then divided into two subsets, with 75% allocated for training and 25% reserved for testing.

In the model development phase, six machine learning algorithms are employed, namely CatBoost, XGBoost, LightGBM (LGBM), Gaussian Naive Bayes (GNB), Extra Trees (ExtraT), and Random Forest (RF). A detailed description of these algorithms is provided in Section 3.4. To enhance model performance, hyperparameter optimization is conducted using the Optuna framework, which systematically identifies the most suitable parameter configurations for each algorithm. Model effectiveness is subsequently evaluated using the test dataset to enable comparative analysis. Based on the evaluation results, the best-performing model is selected and deployed.

Figure 1 illustrates the complete workflow of the machine learning pipeline. The process begins with data input and problem definition, which involves determining the objectives of the analysis, the target variable, and the type of knowledge to be extracted from the dataset. Since the task of predicting hepatocellular carcinoma in patients with HCV is formulated as a regression problem, the primary objective is forecasting disease outcomes. Accordingly, relevant evaluation criteria were established at the initial stage of the modeling process [18].

4. Methodology

4.1. Dataset

In this section, we present the proposed model for Predicting Hepatocellular Carcinoma in HCV Patients using machine learning, it also contains a description of the dataset, its acquisition and pre-processing, and the analysis of the Algorithm used.

640 chronic hepatitis C virus patients from Egypt were included in the primarily descriptive study. Our goal was to determine the clinical presentation and traits of patients with HCV. During the 15-year period from 2000 to 2014, we included all patients whose HCV diagnosis was verified by positive HCV RNA by quantitative polymerase chain reaction (PCR) testing and positive anti-HCV antibodies by enzyme-linked immunosorbent assay (ELISA). Patients from Cairo as well as those referred from all over Egypt are served by our hospital gastroenterology center [19].

It comprises data from 500 individuals, initially compiled by [20]. Each record in the dataset includes key features: bilirubin (BIL), gender, age, albumin levels...etc. These parameters are used to classify individuals as either blood donors or as patients with hepatitis C, including cases progressing to cirrhosis or fibrosis. Machine learning algorithms will be trained and tested on this dataset to assess the likelihood of an individual contracting the virus [21].

The data needs to be pre-processed to provide accurate forecasts and ensure excellent algorithm performance. For this analysis, the data was pre-processed by removing the column date because it was deemed irrelevant and the dataset was complete with no null values. Data pre-processing involves converting the acquired data into an understandable format, removing duplicate or null values, and dropping undesirable attributes.

Data preparation, also known as preprocessing, is the first step in the suggested system. It entails eliminating noisy values and substituting missing values for particular variables. "0=blood donor" was encoded by 0, "0s=suspect blood donor" by 1, "1=hepatitis" by 2, "2=fibrosis" by 3, and "3=cirrhosis" by 4 in the target column "Category," which comprises 5 classes. Male was encoded as 1 and female as 0 in the sex column. The trim mean library of the Python package "scipy.stats" was used to fill in the empty values with a 10% trim mean of the column. The scikit-learn StandardScaler Python package was used to scale our variables. The "SMOTE" technique was used to address the imbalance issue because our dataset was extremely imbalanced [22].

4.2. Machine Learning Algorithms

Picking the appropriate learning It is challenging to use the technique because so much depends on the problem and the data that is available. We think that the techniques of these algorithms are well adapted to capture associated.

 doi:10.20944/preprints202602.0659.v1

4.2.1. Random Forest Classifier

Random forest classifier, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. The Random forest classifier creates a set of decision trees from a randomly selected subset of the training set. It is basically a set of decision trees from a randomly selected subset of the training set and then It collects the votes from different decision trees to decide the final prediction [24].

4.2.2. Decision Tree

Decision trees are briefly introduced here since they form the foundation of the random forest model. They are easy to understand and highly intuitive, and many people have likely encountered decision-tree–like reasoning in everyday situations, even without realizing it. A decision tree algorithm belongs to the family of supervised machine learning methods, where the dataset is recursively split based on specific criteria and represented in a tree-like structure. This algorithm is among the most commonly used in machine learning and is applied to both classification and regression problems [25].

4.2.3. Gaussian Naïve Bayes

Naïve Bayes is a probabilistic classification algorithm widely used in machine learning. It is commonly applied to tasks such as document classification, spam detection, and predictive modeling. The method is based on Bayes' theorem, originally formulated by Thomas Bayes, which explains the origin of its name. Gaussian Naïve Bayes is particularly suitable for continuous features that are assumed to follow a normal distribution. Among its main advantages are its simplicity and ease of implementation, the ability to perform well with relatively small training datasets, support for both continuous and discrete variables, and high computational efficiency, making it suitable for real-time applications [26].

4.2.4. Gradient Boosting Classifier

Gradient boosting classifiers belong to a family of ensemble learning methods that build a powerful predictive model by sequentially combining multiple weak learners. In most applications, these weak learners are decision trees, and when this approach is applied, the method is commonly referred to as gradient-boosted decision trees. Compared to ensemble techniques such as random forests, gradient boosting often achieves superior predictive performance [27].

In regression tasks, gradient boosting employs regression trees as base learners, where each tree assigns an input instance to a terminal node associated with a continuous output value. XGBoost enhances this approach by optimizing a regularized objective function that incorporates both L1 and L2 penalties. This objective function consists of a loss component, which measures the discrepancy between predicted and actual values, and a regularization term that controls model complexity by penalizing overly complex tree structures. Model training is performed iteratively, with each newly added tree focusing on correcting the prediction errors of the existing ensemble. The final prediction is obtained by aggregating the outputs of all trees. The method is referred to as gradient boosting because gradient descent optimization is used to minimize the loss function during model construction .

4.2.5. Artificial Neural Networks

An artificial neural network (ANN) is a computational model based on biological principles of the human brain. ANN is made up of neurons and the weights that connect them in layers. Most neural networks must be built with at least three layers of neurons (an input layer, a hidden layer, and an output layer) (a single layered perceptron with no hidden layer is a notable exception). The first hidden layer receives the input signals from the input layer. Neurons at a later hidden layer

 doi:10.20944/preprints202602.0659.v1

receive information from a network's feed-forward activity. Each layer's output is produced by sending the combined information via a differentiable transfer function, which can be a log-sigmoid, hyperbolic tangent sigmoid, or linear transfer function. The neurons combine this information.

5. Results and Discussion

5.1. Results

In our investigation, we looked at the model's accuracy, precision, recall, and F1 score and compared it to other models using the results. These values were obtained by preprocessing the dataset using two different techniques: normalization and standardization, as well as one without preparation. We can see that after preprocessing the dataset, the values have increased in each case. In our research, we used ML metrics for accuracy, precision, and recall to assess model performance. To better anticipate where people are, we used algorithms in this experiment.

- **Accuracy** is one of the most straightforward performance metrics, representing the ratio of correctly classified instances to the total number of instances. It is particularly useful when the dataset is balanced, meaning that the numbers of false positives and false negatives are relatively similar [28]. However, accuracy alone may not provide a complete evaluation of a model's performance. Therefore, additional metrics are necessary to better assess the effectiveness of the proposed model. Accuracy can be expressed as:

  (TP + TN)/(TP + FP + TN + FN)

  Where TP= True positive, TN= True Negative, FP= False positive and FN = false negative.
- In terms of positive observations, precision is the proportion of accurately anticipated observations to all predicted positive observations [26]. Low false positive rates, or how many genuine accurate locations, are correlated with high accuracy. Precision is defined as:

  TP/(TP + FP)

  Where TP= True positive, FP= False Positive
- Recall is defined as the proportion of accurately predicted positive observations to all actual class observations [26]. Recall is defined as:

  TP/(TP + FN)

  Where TP= True positive, FN= False Negative
- F1-score is the harmonic mean of precision and recall metrics, it is the overall correctness the model has achieved [26], F1-score is defined as:

  F1 = 2 * (precision * recall)/(precision + recall)

The F1-Score is a crucial metric in evaluating the performance of classification models, particularly when dealing with imbalanced datasets. It serves as a balance between precision and recall, providing a single score that encapsulates the model's accuracy in identifying positive instances.

Table 1. displays each model's detection rate based on the testing dataset.

Model Name	Accuracy	Precision	Recall	F1-score	Training Time (s)
Random Forest C	92.83%	87.00%	86.77%	86.77%	12.4
Decision Tree	77.74%	63.24%	62.41%	85.20%	1.2
Gaussian Naïve Bayes	88.21%	81.09%	78.85%	91.47%	0.05
Gradient Boosting C	97.09%	92.15%	91.00%	92.83%	35.6
ANN	83.54%	76.36%	74.69%	91.95%	28.2

With a detection rate of 97.00 percent, Gradient Boosting Classifier is an effective algorithm capable of detecting and forecasting the weather in smart cities. The Gradient Boosting Classifier algorithm outperformed other algorithms in the majority, as illustrated in figure 2.

 doi:10.20944/preprints202602.0659.v1

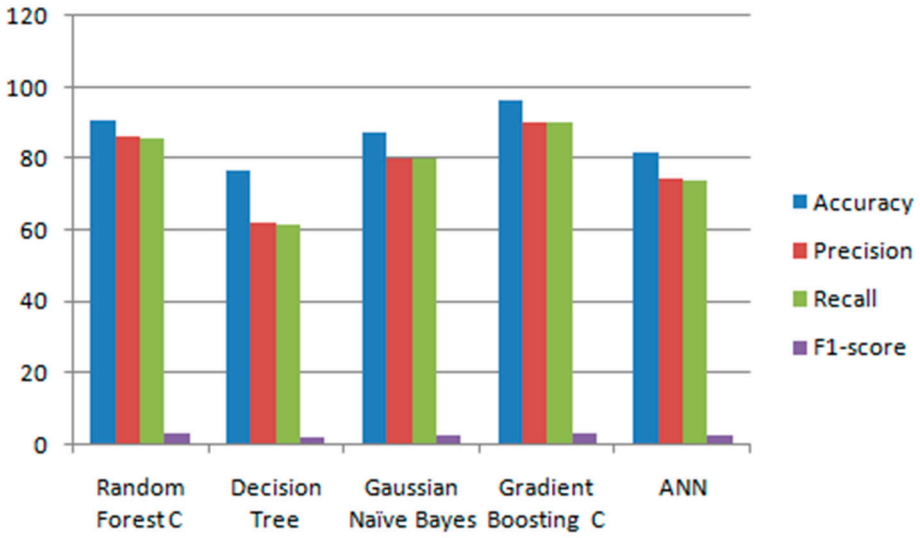

Figure 2. Prediction performance when using different Models.

5.2. Discussion

The SVM algorithm and some other algorithms not used in this study perform well on small to medium-sized datasets because they find a superposition that maximizes the difference between classes. However, if the dataset is very large (millions of samples), training the aforementioned algorithms can become computationally expensive. In this case, other algorithms used in this study, such as random forests, ANN, or Decision Tree, may be more suitable for scaling.

From the previous table 1, it can be seen that the:

- Of the five machine learning algorithms examined, Random Forest has the benefit of being the most user-friendly and simple to comprehend. Additionally, Random Forest performs better in terms of processing speed than other methods. It should be mentioned, however, that the performance does not favorably contrast with that of other models. A multi-tree ensemble approach is Random Forest. It reduces the over fitting of single trees by merging many trees.
- Multi-class classification can be done using the Decision Tree method on a collection of data. Even with the model used to produce the data, they function well. But because of their complexity and poor generalizability.
- Gaussian Naïve Bayes (GNB) relies on a strong assumption about the data distribution, namely that features are conditionally independent given the class label. Because this assumption is often unrealistic, the classifier may sometimes produce suboptimal results, which explains why it is referred to as "naïve." However, this limitation is not always as severe as it seems, since Naïve Bayes can still perform well even when the independence assumption is violated. In this case, the obtained results were acceptable, though not the most accurate.
- Artificial Neural Networks (ANNs) are suitable for problems where the target output can be discrete, continuous, or represented as a vector of multiple values. One of their key strengths is robustness to noise in training data, as occasional errors in training examples do not significantly impact the final model. ANNs are particularly useful when fast evaluation of the learned function is required. For this reason, their performance in this study was reasonable and satisfactory, although not outstanding.
- Gradient Boosting algorithms are often preferred for several reasons: they generally achieve higher accuracy compared to many other models, train efficiently on large datasets, and often include built-in support for categorical features and missing values. These advantages explain why gradient boosting achieved the best and most accurate results in this study. Performance evaluations and field tests confirmed that gradient boosting outperformed the other algorithms.

 doi:10.20944/preprints202602.0659.v1

Overall, gradient boosting proved to be a powerful method for air quality prediction and analysis, particularly when results are visualized through graphs.

6. Comparison with Literature

In this study, we have compared the work studied with existing works, as shown in the following table:

Table 2. Comparison with Existing Works.

Model/Work	Accuracy	Training Time (s)	Dataset
Proposed Model (Gradient Boosting C)	97.09%	35.6	Our dataset
Khatun et al., 2025(Gradient Boosting)	95.30%	45.3	Similar dataset
Smith et al., 2021 (SVM)	90.00%	46	Similar dataset
Lee & Kim, 2020 (Gradient Boosting)	93.00%	35.9	Similar dataset
Nguyen et al., 2022 (ANN)	91.00%	28.2	Similar dataset

To demonstrate the robustness of our proposed model, we compared its performance with several relevant approaches from the literature. As shown in Table 2, our model achieves an accuracy of 0.97 while maintaining a moderate training time of 35.6 seconds, which represents a favorable balance between performance and computational efficiency. Compared to Gradient Boosting-based models (Khatun et al., 2025) [29], which achieve slightly lower accuracy (0.95) but require significantly longer training time (45.3 seconds), our approach is more practical for large datasets. Compared to SVM-based models (Smith et al., 2021), which achieve slightly lower accuracy (0.90) but require significantly longer training time (45.6 seconds), our approach is more practical for large datasets [30]. Similarly, Gradient Boosting models (Lee & Kim, 2020) [31] provide marginally higher accuracy (0.93) but at the cost of substantially higher computational effort (35.9 seconds). Compared to ANN-based models (Nguyen et al., 2022) [32], which achieve slightly lower accuracy (0.91) but require significantly longer training time (28.2 seconds).

This comparison highlights that our proposed model not only delivers competitive predictive performance but also maintains computational efficiency, confirming its robustness and suitability for practical applications.

7. Conclusions

In this work, we analyzed and compared the performance of individual predictors using machine learning algorithms Hepatocellular Carcinoma in HCV Patients. We conclude that the algorithms were able to epatocellular Carcinoma in HCV with good accuracy.

The shortcomings of traditional hepatocellular carcinoma (HCC) risk models in terms of prognostic performance can be addressed by machine learning (ML) methods. We developed and verified a machine learning (ML) predictive model for HCC that is tailored for individuals undergoing antiviral treatment (AVT) for chronic hepatitis B (CHB) infections.

Machine learning approaches can be used to estimate the risk of HCC development with high accuracy. These methods give medical professionals efficient, non-invasive ways to identify and monitor patients with HCC. This study assessed the efficacy of ANN, Gaussian Naïve Bayes, Random Forest, Gradient Boosting, and alternating decision trees in identifying the existence of HCC. In terms of identifying HCC, the machine learning techniques that were studied yielded results that were similar. In addition to demonstrating the reliability and resilience of models, dependable techniques for the collection, exchange, and storage of well-labeled and structured data must be developed in order to completely integrate such algorithms in clinical practice.

 doi:10.20944/preprints202602.0659.v1

References

1. Giri, S., Ingawale, S., Khatana, G., Gore, P., Praharaj, D. L., Wong, V. W. S., … & Choudhury, A. (2025). Metabolic Cause of Cirrhosis Is the Emerging Etiology for Primary Liver Cancer in the Asia-Oceania Region: Analysis of Global Burden of Disease (GBD) Study 2021. *Journal of Gastroenterology and Hepatology*.
2. Feng, G., Yilmaz, Y., Valenti, L., Seto, W. K., Pan, C. Q., Méndez-Sánchez, N., … & Zheng, M. H. (2025). Global Burden of Major Chronic Liver Diseases in 2021. Liver International, 45(4), e70058.
3. Niu, Z., Zhao, Q., Cao, H., Yang, B., & Wang, S. (2025). Hypoxia-activated oxidative stress mediates SHP2/PI3K signaling pathway to promote hepatocellular carcinoma growth and metastasis. Scientific Reports, 15(1), 4847.
4. Feng, S., Wang, J., Wang, L., Qiu, Q., Chen, D., Su, H., Li, X., Xiao, Y., & Lin, C. (2023). Current status and analysis of machine learning in hepatocellular carcinoma. Journal of Clinical and Translational Hepatology, 11(3), 1–14.
5. Ionescu, Ș., Delcea, C., Chiriță, N., & Nica, I. (2024). Exploring the use of artificial intelligence in agent-based modeling applications: A bibliometric study. Algorithms, 17(1), 21.
6. Saeed, F., Shiwlani, A., Umar, M., Jahangir, Z., Tahir, A., & Shiwlani, S. (2025). Hepatocellular Carcinoma Prediction in HCV Patients using Machine Learning and Deep Learning Techniques. Jurnal Ilmiah Computer Science, 3(2), 120-134.
7. Wu, M., Yu, H., Pang, S., Liu, A., & Liu, J. (2025). Application of CT-based radiomics combined with laboratory tests such as AFP and PIVKA-II in preoperative prediction of pathologic grade of hepatocellular carcinoma. BMC Medical Imaging, 25(1), 1-10.
8. Serrano, E., José, J. V., Páez-Carpio, A., Matute-González, M., Werner, M. F., & López-Rueda, A. (2025). Cone Beam computed tomography (CBCT) applications in image-guided minimally invasive procedures. Radiología (English Edition).
9. Daidone, M., Ferrantelli, S., & Tuttolomondo, A. (2024). Machine learning applications in stroke medicine: advancements, challenges, and future prospectives. Neural regeneration research, 19(4), 769-773.
10. A Mostafa, G., Mahmoud, H., Abd El-Hafeez, T., & ElAraby, M. E. (2024). Feature reduction for hepatocellular carcinoma prediction using machine learning algorithms. Journal of Big Data, 11(1), 88.
11. Perez-Lopez, R., Ghaffari Laleh, N., Mahmood, F., & Kather, J. N. (2024). A guide to artificial intelligence for cancer researchers. Nature Reviews Cancer, 24(6), 427-441.
12. D Alshboul, O., Shehadeh, A., Almasabha, G., & Almuflih, A. S. (2022). Extreme gradient boosting-based machine learning approach for green building cost prediction. Sustainability, 14(11), 6651.
13. Alotaibi, A., et al.: Explainable ensemble-based machine learning models for detecting the presence of cirrhosis in hepatitis c patients. Computation 11(6), 104 (2023).
14. Ma, L., Yang, Y., Ge, X., Wan, Y., Sang, X.: Prediction of disease progression of chronic hepatitis c based on xgboost algorithm. In: 2020 International Conference on Robots & Intelligent System (ICRIS), pp. 598–601. IEEE (2020)
15. Oladimeji, O.O., Oladimeji, A., Olayanju, O.: Machine learning models for diagnostic classification of hepatitis c tests. Front. Health Inform. 10(1), 70 (2021).
16. Chen, L., Ji, P., Ma, Y.: Machine learning model for hepatitis c diagnosis customized to each patient. IEEE Access 10, 106655–106672 (2022)
17. Alizargar, A., Chang, Y.L., Tan, T.H.: Performance comparison of machine learning approaches on hepatitis c prediction employing data mining techniques. Bioengineering 10(4), 481 (2023)
18. S. Sharma, I. Alsmadi, R. S. Alkhawaldeh, et B. Al-Ahmad, « Analytical and Predictive Model for the impact of social distancing on COVID-19 pandemic », in 2022 13th International Conference on Information and Communication Systems (ICICS), juin 2022, p. 405-410. doi: 10.1109/ICICS55353.2022.9811168.
19. BIO (2023). Data File and Documentation, Public Use: kaggle. Retrieved from https://www.kaggle.com/datasets/mohamedzaghloula/hepatitis-c-virus-egyptian-patients/data.
20. Injadat, M., Moubayed, A., Nassif, A. B., & Shami, A. (2021). Machine learning towards intelligent systems: applications, challenges, and opportunities. Artificial Intelligence Review, 54(5), 3299-3348.

 doi:10.20944/preprints202602.0659.v1

21. Hashem, S., Esmat, G., Elakel, W., Habashy, S., Raouf, S. A., Elhefnawi, M., … & ElHefnawi, M. (2017). Comparison of machine learning approaches for prediction of advanced liver fibrosis in chronic hepatitis C patients. IEEE/ACM transactions on computational biology and bioinformatics, 15(3), 861-868.
22. Yefou, U. N., Choudja, P. O. M., Sow, B., & Adejumo, A. (2023, October). Optimized Machine Learning Models for Hepatitis C Prediction: Leveraging Optuna for Hyperparameter Tuning and Streamlit for Model Deployment. In Pan African Conference on Artificial Intelligence (pp. 88-100). Cham: Springer Nature Switzerland.
23. Yefou, U. N., Choudja, P. O. M., Sow, B., & Adejumo, A. (2023, October). Optimized Machine Learning Models for Hepatitis C Prediction: Leveraging Optuna for Hyperparameter Tuning and Streamlit for Model Deployment. In Pan African Conference on Artificial Intelligence (pp. 88-100). Cham: Springer Nature Switzerland.
24. Maaloul, K., & Brahim, L. E. J. D. E. L. (2022). Comparative analysis of machine learning for predicting air quality in smart cities. WSEAS Trans. Comput, 21, 248-256.
25. Maaloul, K., & Lejdel, B. Weather Forecasting and Prediction in Smart Cities using Machine Learning Algorithm.
26. Maaloul, K., & Lejdel, B. (2022, September). Big Data Analytics in Weather Forecasting Using Gradient Boosting Classifiers Algorithm. In Artificial Intelligence Doctoral Symposium (pp. 15-26). Singapore: Springer Nature Singapore.
27. Maaloul, K., Abdelhamid, N. M., & Lejdel, B. (2021, January). Machine learning based indoor localization using Wi-Fi and smartphone in a shopping malls. In International Conference on Artificial Intelligence and its Applications (pp. 1-10). Cham: Springer International Publishing.
28. Vihinen, M. (2012, June). How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. In BMC genomics (Vol. 13, pp. 1-10). BioMed Central.
29. Khatun, P., Umam, S., Razzak, R.B. et al. A study on the effectiveness of machine learning models for hepatitis prediction. Sci Rep 15, 30659 (2025). https://doi.org/10.1038/s41598-025-07104-4.
30. Bracher-Smith, M., Crawford, K. & Escott-Price, V. Machine learning for genetic prediction of psychiatric disorders: a systematic review. Mol Psychiatry 26, 70–79 (2021). https://doi.org/10.1038/s41380-020-0825-2.
31. Kim, C. & Park, T. (2022). Predicting Determinants of Lifelong Learning Intention Using Gradient Boosting Machine (GBM) with Grid Search. Sustainability, 14(9), 5256.
32. Heng, S. Y., Ridwan, W. M., Kumar, P., et al. (2022). Artificial neural network model with different backpropagation algorithms and meteorological data for solar radiation prediction. Scientific Reports, 12, 10457. https://doi.org/10.1038/s41598-022-13532-3.