Proceedings of ICMMGH 2026 Symposium: Biomedical Imaging and AI Applications in Neurorehabilitation
DOI: 10.54254/2753-8818/2026.AU30143

Classification of Hepatitis C Progression Using the Hepatitis C Prediction Dataset

Siyi Wang

University of Macau, Macau, China
siyiiigai@gmail.com

Abstract. Hepatitis C virus infection uniquely challenges global healthcare efforts, affecting millions of individuals and often leading to chronic liver disease, including fibrosis and cirrhosis. This study presents a machine learning–based classification framework designed to predict the progression of hepatitis C using a comprehensive patient dataset of demographic and laboratory variables. Several algorithms, including logistic regression, random forest, and XGBoost, were evaluated via cross-validation, with XGBoost demonstrating the highest classification accuracy on the final test set (approximately 93.97%). The models readily distinguish healthy donors from those infected, while also differentiating late-phase disease states such as cirrhosis and fibrosis. Feature engineering, especially selecting a reduced set of predictors (e.g., AST, ALT), proved beneficial for promoting simpler, yet high-performing models with over 93% accuracy. These results underscore the significance of feature selection and careful hyperparameter tuning in improving reliability and interpretability for clinical decision-making. The proposed methodology facilitates better early screening, patient management, and the development of evidence-based strategies tailored to individual patient trajectories, especially for those at risk of disease progression.

Keywords: Hepatitis C, Machine Learning, Classification, Fibrosis, Cirrhosis

1. Introduction

Hepatitis C continues to be one of the leading causes of chronic liver disease worldwide, with an estimated 70 million cases, with approximately 400,000 deaths yearly due to complications from cirrhosis and liver cancer [1]. While early identification and initiation of treatment are the foundation of control and even elimination of this infection, the multicausal pathway of disease limitation can postpone patient assessment, and due to the disease, it can be asymptomatic [2]. The advent of therapies for hepatitis C has improved due to the evolution of therapeutic strategies and is more effective, but distinguishing phases of disease is paramount to planning tailored treatments [3]. Recently, and in part due to machine learning, new methods of using clinical datasets such as the Hepatitis C Prediction Dataset have opened up new possibilities in early case detection and possible disease phase stratification for treatment [4].

The dataset available at the UCI Machine Learning Repository, commonly referenced for evaluating classification performance, contains demographic data (age and sex) and laboratory

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.AU30143

values (albumin, alkaline phosphatase, alanine transaminase, and aspartate transaminase) that enable formal testing of these predictors [5-7]. Thus, the overall task described here is to develop classification models that are accurate and will distinguish healthy patients from Hepatitis C area patients and, furthermore, separate the Hepatitis patients into advancement conditions of cirrhosis and fibrosis [8]. This research is of merit because it may offer clinical decision-makers a model that scales for medically interpreting laboratory-based diagnoses. Also, the research considers the role of feature selection/importance and computational intelligence models to aid in medical diagnostic cases, specifically in Hepatitis C [9].

2. Data exploration and preprocessing

Data exploration and preprocessing are essential steps in the development of a robust and reliable classification model. Prior to implementing any machine learning algorithm, it is important to understand the structure and quality of the dataset, identify and address errors or missing values, and transform the data appropriately. These steps ensure that subsequent modeling yields valid and reliable results.

In the present study, the dataset considered consists of 615 records with 14 attributes. The attributes consist of numerical and categorical variables related to the progression of hepatitis. A preliminary review of the dataset structure using the “str” command reveals a data frame that contains a unique patient ID variable (X), a diagnosis category, demographic variables (Age, Sex), and a series of laboratory values such as Albumin(ALB), Alkaline Phosphatase(ALP), Alanine Aminotransferase (ALT), Aspartate Aminotransferase (AST), Bilirubin (BIL), Cholinesterase (CHE), Cholesterol (CHOL), Creatinine (CREA), Gamma-Glutamyl Transferase (GGT), Protein (PROT). Among these, Category and Sex are represented as character strings, while the remaining features are numeric. A summary analysis of each column was conducted to explore the range and distribution of the data. For instance, patient age ranges from 19 to 77 years of age. Certain laboratory values exhibit a wide range of values, most notably GGT, which spans from 4.50 to 650.90, suggesting potential outliers or variations from clinically relevant extremes.

Missing data is a common issue in clinical datasets. An initial inspection using the "anyNA" function indicated that missing values are present in this dataset. To gain a more comprehensive view of the pattern and extent of missingness across variables, the "vis_miss" function from the naniar package was utilized. This visualization highlighted that while a certain number of variables (e.g., ALB and ALP) had a small number of missing values, others (e.g., BIL and CHOL) had more noticeable gaps. Thus, this paper can address the potential biases that could result from the missing observations with a straightforward data cleaning method consisting of removing all rows with missing values in any variable using the "filter(complete.cases(.))" function. This is a very simple approach to data cleaning, but it is important to recognize that removing rows may affect sample size and statistical power. Depending on the amount of missing data, other more sophisticated approaches such as median/mean imputation or multiple imputation might be a better option.

Following the removal of incomplete rows, categorical variables were converted to appropriate formats for modeling. In the dataset, category is converted to the factor type, which allows the algorithms to differentiate among diagnostic classes. Likewise, sex is recast as a factor in the model with “m” and “f” as the two levels. The X variable, which represents patient ID, is omitted from any further analysis since it is not directly relevant to disease classification tasks.

To further assess the relationships among numeric features and detect possible multicollinearity, a correlation matrix was calculated. As shown in Figure 1, several strong and moderate correlations were observed. For example, ALP and GGT exhibited a relatively strong positive correlation, while

DOI: 10.54254/2753-8818/2026.AU30143

ALB showed moderate positive correlations with both CHE and PROT. The strength of these relationships will guide feature selection decisions, or possibly even further data transformation. Regularization or dimension reduction approaches will help to address high collinearity in order to avoid inflating the variance of estimated parameters.

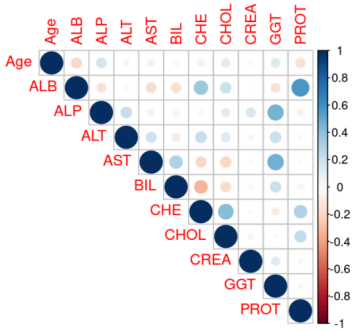


Figure 1. Correlation matrix of numeric predictors

To sum up, the exploration and preprocessing of the hepatitis C dataset are guided by three key considerations: addressing missing data (and deciding on its imputation method), encoding categorical variables, and analyzing the distributions and correlations of variables. Taking the time to ensure data is clean and logically consistent at this stage will make subsequent modeling efforts—including feature selection and classification—more likely to generate accurate and reliable predictions. As a component of data exploration, the author stresses that systematic handling of data, together with clear reasoning behind data transformations, adjustments, or imputation, is fundamental to producing valid and interpretable results for clinical research.

3. Machine learning methodologies

3.1. Model selection and training

The main goal of this step is to select the best classification algorithm in order to predict the progression of hepatitis C. After data exploration and cleaning, the dataset was split into training and test parts so that we could ensure later evaluations of model performance would not be biased with respect to the training process. The caret function “createDataPartition” split 80% of the cleaned dataset (train_data) for model development and the remaining 20% (test_data) for final assessments of model validation.

To ensure compatibility with machine learning algorithms, categorical factor levels were renamed to remove spaces and special characters (e.g., “0=Blood Donor” was modified to “BloodDonor”). This process was necessary, as some machine learning algorithms create probability columns internally and cannot work with invalid R variable names. Therefore, ensuring factor levels are syntactically valid minimizes issues when class-specific probability estimates are computed. To improve robustness and generalizability, a repeated cross-validation scheme was implemented. The trainControl function allowed for five-fold repeated cross-validation (repeats = 1) to partition the training data into five folds, rotating through the five folds as the validation subset sequentially. Averaging model performance metrics across all folds helps to prevent the likelihood of overfitting while generating relatively stable model accuracy estimators. This is also encouraged through the process of using hyperparameter tuning based on the development of reliable performance assessments. Furthermore, many advanced texts on machine learning methods discuss the

DOI: 10.54254/2753-8818/2026.AU30143

importance of rigorous cross-validation combined with systematic exploration of hyperparameters in order to maximize the chance of final model generalizability on varying clinical populations.

To represent the various methodological strategies, three classification algorithms were chosen:

Logistic Regression (multinomial): The caret's "multinom" method takes logistic regression into a multi-class setting, wherein the class memberships are treated as log-odds. This method is interpretable and depends on linear decision boundaries.

Random Forest: This is an ensemble approach that combines many decision trees, where each decision tree is trained on a bootstrap sample of the data set and the predictions are made by taking a majority vote. Random forests are typically known to deal well with complex interactions and heterogeneous data and usually contain a measure of variable importance [7].

XGBoost: This method, the gradient boosting trees algorithm, builds trees in a sequential manner where each tree-building attempt tries to correct the residual error of the model before it. In multiple comparisons of alternative approaches, this method is often the most performant with tuning of hyperparameter options on tabular data sets (for example, learning rate, maximum depth of trees, subsampling ratios, etc) [10]l

The three methods are chosen for their complementary strengths matching the study's needs. Multinomial Logistic regression offers strong interpretability and relies on linear decision boundaries, serving as an interpretable baseline. Random Forest, an ensemble method, handles complex variable interactions and heterogeneous data well while providing feature importance insights. XGBoost, a gradient-boosting algorithm, excels at tabular data by correcting residual errors sequentially, often delivering top performance with proper hyperparameter tuning—critical for accurately classifying hepatitis C stages.

For training of each model, all hyperparameter settings were evaluated using the caret tuneLength parameter. As an illustration, for multinomial logistic regression, the search examined different values of the regularization parameter (decay), while for the random forest, the model explored bandwidth (mtry) values to allow complexity of each tree. The training loop of the XGBoost evaluated nrounds, max_depth, eta, gamma, colsample_bytree, min_child_weight, and subsample parameters. The best-performing combination of the hyperparameter settings was identified by composite scores from cross-validation folds for each individual algorithm.

After the repeated cross-validation phase was complete, the resamples function was carried out to compare the overall performance of several algorithms (GLM, RF, and XGB). Important statistics (mean accuracy, AUC, Kappa, and log loss) were summarized. Together these statistics provide a rounded perspective of the predictive performance, class separation quality, and calibration of the model. The summary shows that all three models obtained excellent classification performance (mean accuracy greater than 0.94).

Random Forest and XGBoost both perform better than multinomial logistic regression regarding classification accuracy and AUC. XGBoost performs the best of all models with the highest mean accuracy, as the iterative, boosting-based method facilitates performance across evaluations [10]. The simplest, most interpretable model, logistic regression, performs relatively close to the leadership models. In review, these results also demonstrate the need for model experimentation and systematic evaluation to explore approaches that allow for the highest predictive reliability for progression into hepatitis C.

Table 1 is a summary table of the models' mean performance metrics from five-fold cross-validation, along with their identified best tuning parameters:

DOI: 10.54254/2753-8818/2026.AU30143

Table 1. Comparison of model performance and best tuning parameters

Model	Best Tuning	Mean Accuracy	Mean AUC	Mean Kappa	Mean logLoss
GLM	decay = 0.1	0.9429	0.9069	0.7191	0.6801
RF	mtry = 7	0.9535	0.9864	0.7386	0.138
XGBoost	nrounds=50, max_depth=2, eta=0.3	0.9598	0.9841	0.7777	0.136

3.2. Feature engineering techniques

Feature engineering is an essential step in the process of improving a machine learning workflow, particularly with high-dimensional data or correlated predictors. For this project, we employed a targeted approach to feature selection to identify the most informative subset of laboratory measurements in order to increase the predictive performance of the model. In our case, we leveraged the Recursive Feature Elimination (RFE) procedure with random forests, which is an efficient procedure to identify the variables that maximize classification accuracy and discard the variables that add noise or redundancy.

The RFE process was set up to evaluate several candidate feature subset sizes using cross-validation (with three folds). Random forests were selected as the underlying algorithm because of their strength for handling complex interactions and feature-importance insights. Along each selection round, the performance of each subset was estimated on accuracy and Kappa, an agreement index, reflected chance-level classifications. A subset size of two variables displayed the highest average accuracy and Kappa, with AST and ALT selected as the most appropriate variable predictors for distinguishing diagnostic subsets.

After identifying the two most important variables, a model that included only AST and ALT was built in the caret train function using the random forest algorithm. Benefits of selecting a subset of the original features include decreasing model complexity, increasing interpretability, and potentially increasing generalizability [11, 12]. While complicated transformations or reviews of interactions would also be meaningful, RFE for feature selection may be an appropriate first step toward increasing the predictive power of clinical datasets and could lead to improvement in the precision and interpretability of models.

4. Results and discussion

4.1. Comparative performance analysis

When evaluating the performance of the final models, classification accuracy and confusion matrices on the held-out test sets are key metrics. Out of the models considered—logistic regression (GLM), random forest, XGboost, and the random forest arranged using only the two top features—XGboost produced the greatest accuracy, at nearly 93.97% accuracy, followed closely behind by the estimator of the full-feature random forest and the random forest model with only the two relevant features, both at approximately 93.10%, as stated in Panel B of Figure 2. Although the logistic regression performed comparably (93.10%), we noted a slightly lower value for Kappa, a metric of agreement accounting for chance. The balanced accuracy metrics for XGBoost demonstrate good reliability and performance of detection for minority class detection, and, from an ordinal perspective, for SuspectDonor, Fibrosis, and Cirrhosis.

DOI: 10.54254/2753-8818/2026.AU30143

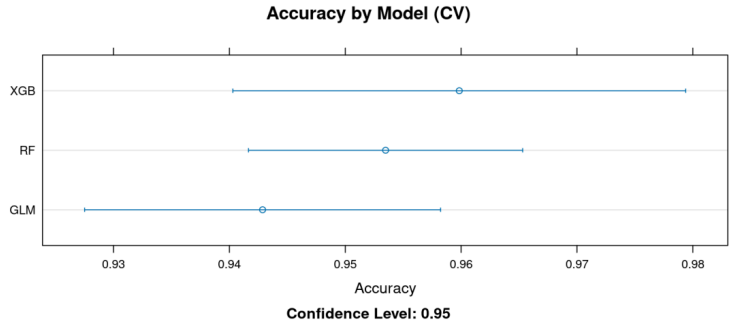


Figure 2. Dotplots for comparing models on accuracy

BloodDonor class due to the high frequency of samples in the dataset. However, several classes demonstrated low An examination of sensitivity and specificity indicates that most of the models perform reasonably well for the sensitivity for less frequent categories –and in particular the SuspectDonor and Hepatitis categories were again more difficult for the models to classify. XGBoost more easily gains a state of balance in sensitivity and specificity, with some classes still indicating a difficulty in classifications that could potentially benefit from engineered features or larger samples in the training data.

4.2. Implications for clinical decision-making

From a clinical decision-making standpoint, these findings underscore the value of ensemble methods in identifying key progression stages of hepatitis C. Both the XGBoost and random forest models can accurately stratify patient risk or determine disease progression using a broad range of laboratory measurements. What is more encouraging is that even the random forest model trained solely on AST and ALT yielded strong performance—this could translate to cost reductions or simplified patient assessment processes, especially in settings with limited resources. Ultimately, model interpretability remains crucial: clinicians need to go beyond grasping the model's predictive accuracy; they must also understand how to apply this information to better evaluate individual patients and implement earlier interventions.

5. Conclusion

This study aimed to develop a machine learning-based classification framework for predicting hepatitis C progression to improve disease diagnosis and management. It utilized the Hepatitis C Prediction Dataset from the UCI Machine Learning Repository, which contained 615 records with 14 attributes, including demographic data (age, sex) and laboratory values (e.g., ALB, ALT, AST).

In the preprocessing phase, rows with missing values were removed, categorical variables (e.g., diagnosis category, Sex) were converted to factor types, and variable correlations were analyzed to address multicollinearity. The dataset was split into an 80% training set and a 20% test set. Three algorithms—logistic regression (multinomial), random forest, and XGBoost—were evaluated, with their hyperparameters optimized via 5-fold cross-validation.

For feature engineering, Recursive Feature Elimination (RFE) combined with random forest was used, identifying AST and ALT as the key predictive indicators. On the test set, XGBoost achieved the highest accuracy (approximately 93.97%), while the random forest model using only AST and ALT, as well as the logistic regression model, both reached an accuracy of around 93.10%.

DOI: 10.54254/2753-8818/2026.AU30143

The study demonstrated that machine learning can effectively distinguish different stages of hepatitis C, with XGBoost performing best. The model with simplified indicators balances accuracy and cost-effectiveness, supporting clinical early screening and personalized treatment. Future research may expand the sample size and explore deep learning to enhance model performance.

References

[1] Hoofnagle, J. H. (2002) Course and outcome of hepatitis C. Hepatology, 36(S1): S21-S29.
[2] Scott, J. D., & Gretch, D. R. (2007) Molecular diagnostics of hepatitis C virus infection: a systematic review. Jama, 297(7): 724-732.
[3] Obaido, G., Ogbuokiri, B., Swart, T. G., Ayawei, N., Kasongo, S. M., Aruleba, K., ... & Esenogho, E. (2022) An interpretable machine learning approach for hepatitis b diagnosis. Applied sciences, 12(21): 11127.
[4] Mahmoud, A., Awad, N. A., Alsubaie, N., Ansarullah, S. I., Alqahtani, M. S., Abbas, M., ... & Saber, A. (2023) Advanced deep learning approaches for accurate brain tumor classification in medical imaging. Symmetry, 15(3): 571.
[5] Lichtinghagen, R., Klawonn, F., & Hoffmann, G. (2020) HCV data [Dataset]. UCI Machine Learning Repository. https: //doi.org/10.24432/C5D612.
[6] Gorunescu, F., & Gorunescu, F. (2011) Classification performance evaluation. Data Mining: Concepts, models and techniques: 319-330.
[7] Badawy, M., Ramadan, N., & Hefny, H. A. (2023). Healthcare predictive analytics using machine learning and deep learning techniques: a survey. Journal of Electrical Systems and Information Technology, 10(1), 40.
[8] Meshram, R. J., Kathwate, G. H., & Gacche, R. N. (2022) Progress, evolving therapeutic/diagnostic approaches, and challenges in the management of hepatitis C virus infections. Archives of Virology, 167(3): 717-736.
[9] Sheet, F. (2015) Hepatitis E. In. World Health Organization, available at http: //www. who. int/mediacentre/factsheets/fs280/en/accessed, 31.
[10] Friedman, J. H. (2001) Greedy function approximation: a gradient boosting machine. Annals of statistics: 1189-1232.
[11] Guo, A.K. (2022) Advances in Applied Mathematics, 11: 3018.
[12] Yang, L.Y. & Wen. G. (2020) Application of deep learning in medical imaging research. Journal of Molecular Imaging, 43(2): 183r187.