InfoTech Spectrum: Iraqi Journal of Data Science (IJDS)
ISSN: 3007-5467, Vol 3(No.2), 44-55 (2026)
DOI: https://doi.org/10.51173/ijds.v3i2.82
https://ijds.mtu.edu.iq

HEPATITIS C DISEASE PREDICTION USING MACHINE LEARNING MODELS WITH SMOTE-BASED DATA BALANCING

Omar Shakir Hasan*

Directorate of Education, Nineveh, Mosul, Iraq

** Corresponding author E-mail: omarshakir06@gmail.com* **(Omar Shakir Hasan)**

RESEARCH ARTICLE

ARTICLE INFORMATION

SUBMISSION HISTORY:
Received: 2 May 2026
Revised: 15 June 2026
Accepted: 27 June 2026
Published: 30 June 2026

KEYWORDS:
Hepatitis C;
Machine Learning;
Ensemble Learning;
XGBoost;
SMOTE.

ABSTRACT

The problem of hepatitis C is worldwide, with many potential health issues associated with this serious disease (e.g., hepatitis C may lead to liver cirrhosis in some patients). Accurate and early diagnosis of Hepatitis C will improve patient care and possibly lower the death rate from this virus. This paper examined the use of machine learning models to classify patients' hepatitis C stage using laboratory-based clinical data from the hepatitis C dataset on Kaggle, with n=615 subjects. The data was split with 80% for training and 20% for testing. A preprocessing methodology was implemented on the dataset to obtain the best fit for multiple models. For example, missing data were imputed using KNN; categorical variables were encoded as numbers (label encoding); and each feature was scaled (standard scaler). To address class imbalance, we used the Synthetic Minority Over-sampling Technique (SMOTE). The models we implemented were: XGBoost, Random Forest, Decision Tree, Logistic Regression, and Gaussian Naïve Bayes. Our results show that the XGBoost/Synthetic Minority Over-sampling Technique performed best, achieving 95% classification accuracy and a macro F1 score of 0.95, compared to the other models we tested. The results provide additional evidence that combining SMOTE with ensemble learning offers a robust decision-support solution for classifying patients with Hepatitis C in a multiclass setting and an efficient means for early clinical diagnosis.

1. INTRODUCTION

Hepatitis C is a hepatitis virus (Hepatitis C virus (HCV)) infection of the liver, and is a significant public health problem because of its high prevalence and potential for numerous long-term complications [1]. Often, the disease lies dormant, and many patients do not have symptoms in the early stages. This means that many patients are diagnosed with advanced liver damage, such as fibrosis, cirrhosis, and hepatocellular carcinoma[2]. The late diagnosis also limits the effectiveness of treatment and increases mortality, underscoring the need for more accurate and prompt diagnostic methods [3]. The most common diagnostic approaches for Hepatitis C are biochemical and serological blood tests in the laboratory, including liver enzyme levels and viral RNA[4]. The methods used are clinically reliable but can be time-consuming, resource-intensive, and sometimes inadequate to represent complex relationships among many clinical variables [5]. In addition, even when data like this exists, it can only be interpreted by an expert, and that may not always be available, especially in resource-limited settings[6]. As a result, there is increased interest in developing intelligent systems that can help clinicians make decisions and render diagnoses more efficiently. Over the years, machine learning (ML) has become a game-changer in the health care industry, enabling advanced predictive modeling and pattern recognition [7].

ML algorithms can analyze vast amounts of clinical and biochemical data to identify patterns and relationships that are difficult to detect with traditional statistical methods [8]. While machine learning shows promise for medical data, several challenges arise in building reliable predictive

DOI: https://doi.org/10.51173/ijds.v3i2.82

models. In practice, clinical datasets can be incomplete, noisy, and imbalanced, leading to problems in modeling and generalization [9]. For instance, in the context of disease progression, there may be fewer observations at certain disease stages, leading models to predict only for the majority class [9]. Thus, appropriate data preprocessing methods (imputation, normalization, resampling) are essential for developing unbiased and valid models. [10]

There has been considerable progress made in past studies; however, limitations do still exist. Most existing studies are binary classification studies that do not use balanced datasets. Additionally, they generally evaluate only a small number of algorithms (machine learning models) and focus solely on accuracy, without adequately assessing the impact of combining all preprocessing strategies — KNN imputation, feature scaling, and SMOTE — on predictive accuracy. Furthermore, for multiclass classification of Hepatitis C, there are too few shared comparisons between classical statistical techniques and ensemble machine learning approaches. As such, there remains a need for an integrated framework that combines appropriate preprocessing methods with strong ML estimation/classification techniques to achieve optimal multiclass prediction performance. This paper addresses these issues by designing a machine learning model for classifying Hepatitis C disease stages. This framework combines several preprocessing methods, including handling missing data with K-Nearest Neighbors (KNN) imputation, balancing class distributions with Synthetic Minority Over-sampling Technique (SMOTE), and feature scaling for data normalization. Moreover, various machine learning approaches, from traditional models to ensemble techniques, are used to assess and compare predictive performance in this study, creating an accurate, reliable, and scalable predictive system that could assist clinical decision-making and improve early disease detection. The main contributions of this study are as follows:

- An integrated machine learning framework for multiclass Hepatitis C classification is presented.
- KNN imputation, feature scaling, and SMOTE are combined during data preprocessing.
- The performance of classical and ensemble machine learning models is compared using multiple evaluation metrics.

2. RELATED WORK

In[11], Ali et al. developed an explainable machine learning (XAI) framework for diagnosing hepatitis C (HCV) using Sequential Forward Selection (SFS) and SHapley Additive exPlanations (SHAP) to enhance the interpretability of predictions. A real-world dataset taken from Jordan University Hospital was used, including 1801 patients and 13 clinical features. In this study, several machine learning algorithms were tested across different scenarios, including Logistic Regression (LR), K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), and Neural Networks (NN), both with and without the Synthetic Minority Oversampling Technique (SMOTE). The experimental results indicate that the accuracy of KNN is 83.1%, and it is the best overall performance. In [12], Edeh et al. proposed an ensemble learning approach based on artificial intelligence to predict Hepatitis C disease from clinical and laboratory parameters. The study integrated an MLP, a Bayesian Network, and a QUEST decision tree to enhance classification accuracy compared to the individual models. The results revealed that each of the individual models obtained 94.10%, 94.47%, and 94.63%, respectively, in the accuracy of the classification based on the use of the publicly available dataset without any preprocessing, such as removing the missing values and outliers from the dataset. The best ensemble model achieved the highest accuracy of 95.59%. The authors concluded that ensemble learning is more effective at detecting Hepatitis C and advanced liver fibrosis than single classifiers, as it further enhances predictive accuracy, reduces model bias, and improves generalization.

In[13], Khatun et al. suggested a machine learning model for the prognosis of survival outcomes in hepatitis based on the UCI hepatitis data set containing 155 patients and 20 clinical features. With the help of Boruta, this study aimed to find important predictors such as Ascites, Varices, Bilirubin, Age, Spiders, and Alkaline Phosphate. Different classification algorithms were tested, including Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Artificial Neural Network (ANN), Random Forest (RF), AdaBoost, and XGBoost. The

experiments showed that the best-performing model was the Random Forest model, with 92.42% accuracy, 96.77% precision, 95.24% sensitivity, and 96.00% F1-score. The study has demonstrated that incorporating the Boruta-based feature selection method with ensemble learning approaches can improve hepatitis prognosis prediction and support early diagnosis and healthcare decision-making. The study in [14] aimed to use machine learning (ML) and explainable AI (XAI) to enhance the diagnosis of Hepatitis C Virus (HCV) based on 615 patient records. The accuracy of the Random Forest (RF) and XGBoost models was 93.75% and 92.38%, respectively. To increase interpretability, important clinical biomarkers, including ALT, AST, bilirubin, albumin, and age, were selected and analyzed using SHAP and ALE methods. The findings indicated that the accuracy of prediction and the transparency of the models were significantly improved when using both ML and XAI, facilitating rapid, accurate, and interpretable HCV screening and ultimately aiding early diagnosis and improved clinical decision-making. In [15], the suggested study presents an Adaptive Preprocessing Technique that uses Mean Imputation, Outlier Removal, Log Normalization, Feature Selection, Feature Scaling, and Data Balancing techniques. Moreover, several ensemble models were created by combining basic ML algorithms, and subsequently, hyperparameter optimization was conducted. The performance of the proposed hyper-tuned ensemble Logistic Regression (LR) model was excellent, achieving 99.87% training and 99.80% testing accuracy. In [16], a proposed framework for early detection of hepatitis C virus (HCV) was presented. The study introduced a Cascade RF–LR model that combines Random Forest (RF) and Logistic Regression (LR) algorithms to improve multiclass HCV classification, particularly for imbalanced clinical datasets. The Synthetic Minority Oversampling Technique (SMOTE) was applied during preprocessing, while the Artificial Bee Colony (ABC) algorithm was utilized to optimize the threshold between the two classification stages. The model achieved strong results in terms of accuracy, Precision, Recall, F1-score, and Matthews Correlation Coefficient (MCC). The study in [17] compared the performance of different machine learning algorithms to predict Hepatitis C virus using the best accuracy for binary classification; Random Forest outperformed all the other algorithms with 75.05%, and for multiclass classification, SVM had the best performance with 28.45%. The results showed that binary classification performed better than multiclass classification. The study also emphasized the significance of feature selection, normalization, and data scaling in enhancing the model's accuracy and efficiency.

Further, to ensure the interpretability of the model, Bayesian Networks and SHAP were employed to explain the factors that lead to the model's predictions, highlighting the importance of explainable AI in healthcare applications. The proposed study in[[18]] aimed to predict the presence of Hepatitis C Virus (HCV) in real-world medical datasets with machine learning, with a particular focus on the problem of highly imbalanced datasets. To enhance performance, it used feature selection techniques like RFE, ANOVA, and correlation-based selection, and balancing techniques like SMOTE, Borderline-SMOTE, SVMSMOTE, and ADASYN. Random Forest, Decision Tree, XGBoost, AdaBoost, and Logistic Regression were used for several classifiers. The results demonstrated the feasibility of using ensemble methods; the highest Recall was obtained (0.86) by AdaBoost with ADASYN. In general, it was found that using ensemble learning with the effective sampling methods greatly improves the prediction of HCV and facilitates more accurate clinical decision-making. In [19], machine learning techniques have been employed to perform early detection and diagnosis of Hepatitis C Virus. A model was created using the UCI Hepatitis C dataset to classify five outcomes: (1) Blood Donor (2) Hepatitis (3) Fibrosis (4) Cirrhosis (5) suspected blood donor. The Random Forest classifier showed that the best results are obtained for multiclass prediction, with an accuracy of 93.5%, which could be used for early diagnosis and clinical decision-making. Sharma et al. [20] proposed a machine learning framework for Hepatitis C detection using a cross-dataset meta-model with a multi-dimensional pre-clustering approach. The study combined two publicly available datasets and evaluated several machine learning models, including XGBoost, Random Forest, K-Nearest Neighbors, and Support Vector Classifier. In addition, explainable artificial intelligence techniques were employed to improve model interpretation. The proposed meta-model achieved a prediction accuracy of 94.82%, slightly outperforming the baseline Random Forest model. The authors concluded that combining multiple datasets and pre-clustering can improve prediction performance and model generalization. Sunori et al. [21] developed a machine

learning technique for predicting multiclass outcomes of Hepatitis C using clinical lab tests in conjunction with PCA (principal component analysis). They successfully reduced the number of features for 615 patients in the HCV study and applied multiple machine learning classifiers to the data, thereby supporting effective multiclass classification of hepatitis C. The literature review has shown that many aspects of predicting Hepatitis C have been studied, including explainable AI/ensemble learning/feature selection. However, limited research has compared the performance of classical and ensemble machine learning algorithms for multiclass Hepatitis C classification, using a common preprocessing approach combining KNN imputation, feature scaling, and SMOTE.

3. METHODS

The proposed methodology for Hepatitis C classification outlines the main stages of the paper's framework, including data preprocessing. It incorporates imbalance handling techniques and machine learning classification models to improve predictive performance. Fig. 1 presents the overall proposed methodology and the sequence of processes followed in this study.

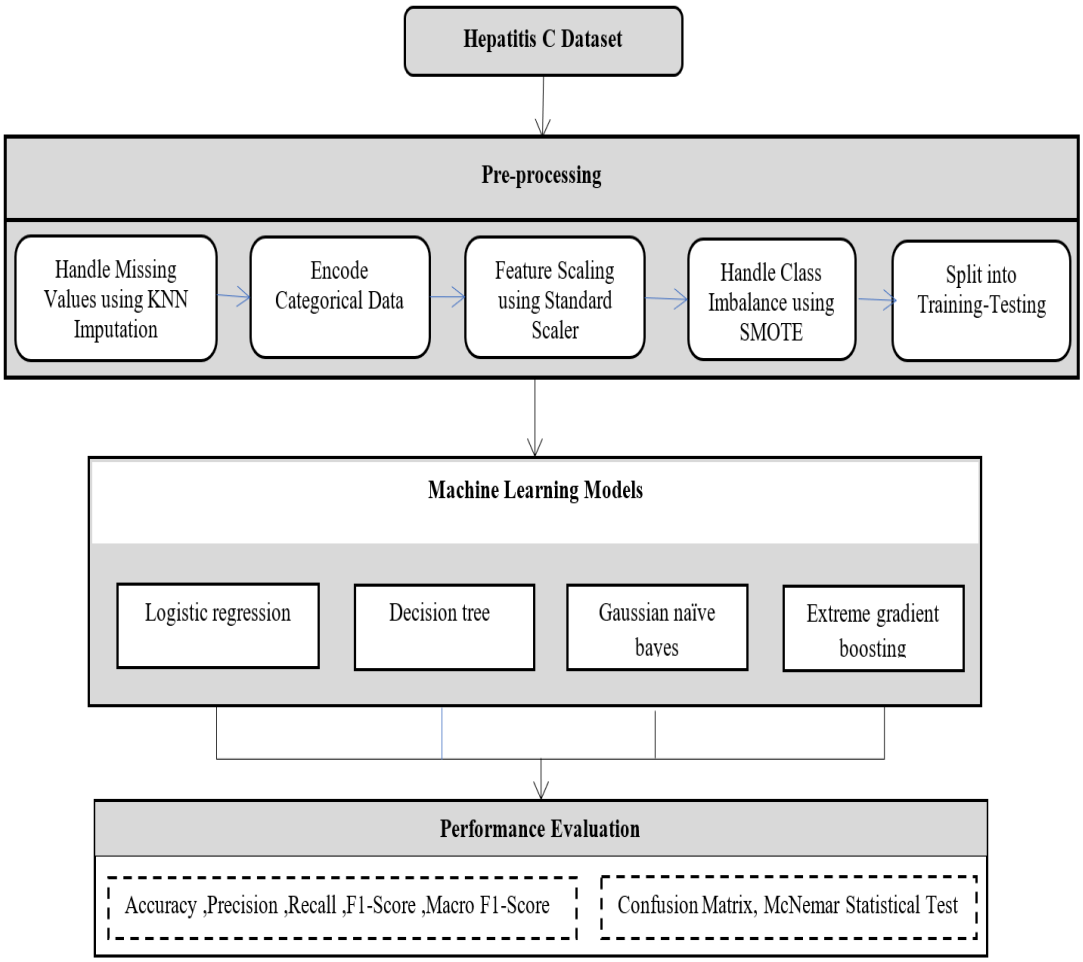

Figure 1. Proposed Methodology

3.1. Dataset Description

The dataset for this research includes laboratory and clinical parameters related to hepatitis C diagnosis. Basic demographic information, as well as a patient's age and sex, provides the basis for some contextual characteristics. The dataset also contains several laboratory values (ALB, ALP, ALT, AST, BIL, CHE, CHOL, CRE, GGT, PROT) related to liver function and other associated biochemical processes. The target variable in this study is a multiclass label indicating different stages of hepatitis C disease [22][23]. Table 1 includes the major parameters from the hepatitis C dataset used in this research study.

Table 1. Characteristics of the Hepatitis C Dataset

Characteristic	Description
Dataset source	UCI Machine Learning Repository (Hepatitis C Dataset)
Number of instances	615
Number of features	12 input features + 1 target variable
Feature types	10 numerical features, 2 categorical variables (Age and laboratory biomarkers are numerical; Sex and Category are categorical)
Target classes	Blood Donor, Suspect Blood Donor, Hepatitis, Fibrosis, Cirrhosis
Missing values	Present in several laboratory variables
Laboratory biomarkers	ALB, ALP, ALT, AST, BIL, CHE, CHOL, CREA, GGT, PROT

3.2.Data Preprocessing

To guarantee quality and enhance performance, data preprocessing has been performed. Label encoding has been applied to categorical features (e.g., sex and category, the target variable) by converting them to numeric values. KNN imputation has been used to handle missing values with k = 5. Each missing value will be estimated from the average of the five nearest neighbors. Standardization was applied using StandardScaler, which set the data to zero mean and unit variance for uniformity across features of different scales. It was also done to ensure there were no class imbalances in the training set by using the Synthetic Minority Over-sampling Technique (SMOTE), which generated synthetic samples for minority classes to enhance the model's generalization. The dataset was divided into training and testing sets using an 80:20 stratified split. A random seed of 42 was used to ensure the reproducibility of the experimental results. The hyperparameter settings used for the evaluated machine learning models are summarized in Table 2. Unless otherwise specified, the remaining hyperparameters were kept at their default values.

Table 2. Hyperparameter settings of the evaluated machine learning models

Model	Hyperparameter settings
Logistic Regression	Solver = lbfgs, Max iterations = 1000, Random state = 42
Decision Tree	Criterion = gini, Random state = 42
Random Forest	n_estimators = 100, Criterion = gini, Random state = 42
Gaussian Naïve Bayes	Default parameters
XGBoost	n_estimators = 100, Learning rate = 0.3, Max depth = 6, Random state = 42

All remaining hyperparameters were kept at their default values.

3.3 Machine Learning Models

In this study, a series of classical and ensemble machine learning models were evaluated, such as Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Gaussian Naïve Bayes (GNB), and Extreme Gradient Boosting (XGBoost). These models were chosen to give a wide range of comparisons of learning mechanisms and predictive performance. Logistic Regression is a linear baseline model, and Decision Tree is an interpretable rule-based classification model. Random Forest and XGBoost are ensembles of learners, which minimize overfitting. Gaussian Naïve Bayes is a probabilistic model suitable for continuous data. This combination enables a thorough analysis of both simple and advanced methods for classifying Hepatitis C.

3.4. Evaluation Metrics

To evaluate the performance of the classification models, we used several standard metrics, including accuracy, Precision, Recall, and F1-score [24][25].

- Accuracy: the number of correctly classified samples/Total number of samples. It is an

efficient indicator of overall performance but can be misleading when class imbalance is present.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{Total}} \quad \ldots (1)$$

- True Positive (TP): Correctly predicted positive cases (when the actual case is positive).
- TN (True Negative): Negative cases that were correctly predicted.
- TP: True Positive (Correct positive): number of samples with positive results that are actually positive
- Precision: is the percentage of the correctly predicted positive cases that have been predicted as positive. It evaluates the model's ability to avoid false positive errors.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad \ldots (2)$$

- Recall: Recall (sensitivity) is the number of positive instances correctly identified by the model. A high recall value means that there are fewer false negatives.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad \ldots (3)$$

- F1-Score: It is the harmonic average of Precision and Recall. It provides a balanced measure that accounts for false positives and false negatives.

$$\text{F1} - \text{score} = \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \ldots (4)$$

- Recall and F1-score are used as measures of the model's effectiveness in this case because it is a classification problem with imbalanced classes.

In addition to the class-wise evaluation metrics, Macro Precision, Macro Recall, and Macro F1-score were calculated. These metrics are obtained by computing the arithmetic mean of the corresponding metric across all classes, assigning equal importance to each class regardless of its size. Macro-averaged metrics provide a balanced evaluation for multiclass and imbalanced classification problems.

$$\text{Macro Precision} = \frac{1}{C}\sum_{i=1}^{C} Precisio\, n_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(5)$$

$$\text{Macro Recall} = \frac{1}{C}\sum_{i=1}^{C} Recal\, l_i \ldots\ldots\ldots\ldots\ldots\ldots(6)$$

$$\text{Macro F1} = \frac{1}{C}\sum_{i=1}^{C} F\,1_i \ldots\ldots\ldots\ldots\ldots(7)$$

Where: C is the number of classes.

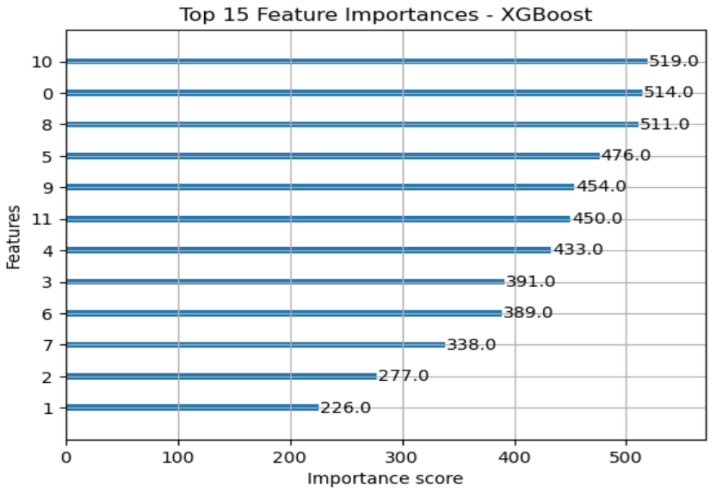

Figure 2. Feature importance

3.5.Feature Importance Analysis

Feature importance analysis was conducted on the XGBoost model to elucidate the importance of each feature in the classification process. The most important features are shown in Fig. 2, ranked by importance. The results indicate that a few features were most important for the model's performance; namely, those with importance scores greater than 500. Overall, the distribution suggests that XGBoost heavily weights the most important features and less so the less important ones. This underlines the model's capacity to zero in on the most pertinent clinical features.

3.6 Experimental Design and Reproducibility

The experiments were conducted using Python 3.11 with the Scikit-learn and XGBoost libraries on a personal computer running Windows 11. A stratified 80:20 train-test split was adopted to preserve the class distribution in both subsets. A random seed of 42 was used to ensure reproducibility, and SMOTE was applied only to the training data to prevent data leakage.

4. RESULTS

4.1. Model Performance Without SMOTE.

Table 3 shows the performance of five machine learning models on the Hepatitis C dataset before applying SMOTE to the Hepatitis C dataset.

Table 3. Performance of machine learning models before SMOTE.

Model	Class	Precision	Recall	F1-Score	Accuracy
Gaussian Naïve Bayes	0	0.97	0.98	0.98	0.93
	1	0.33	1.00	0.50	
	2	0.67	0.40	0.50	
	3	0.50	0.50	0.50	
	4	1.00	0.83	0.91	
Logistic Regression	0	0.97	1.00	0.99	0.96
	1	0.09	0.08	0.06	
	2	0.75	0.60	0.67	
	3	0.75	0.75	0.75	
	4	1.00	0.83	0.91	
Decision Tree	0	0.95	0.98	0.97	0.90
	1	0.15	0.16	0.15	
	2	0.20	0.20	0.20	
	3	0.50	0.25	0.33	
	4	0.80	0.67	0.73	
Random Forest	0	0.96	1.00	0.98	0.94
	1	0.17	0.15	0.12	
	2	0.50	0.20	0.29	
	3	0.75	0.75	0.75	
	4	1.00	0.83	0.91	
XGBoost	0	0.97	1.00	0.99	0.95
	1	0.18	0.19	0.16	
	2	0.60	0.60	0.60	
	3	0.67	0.50	0.57	
	4	1.00	0.83	0.91	

The results show a difference in model performance for various classes. All models had good performance with the majority class (Class 0), but inconsistent and weak performance with minority classes (Class 1-4). High accuracy for Class 0, but low Precision, Recall, and F1 scores for minority classes in all the models, especially Class 1 and Class 2 – Gaussian Naïve Bayes, Logistic Regression, Decision Tree, Random Forest, and XGBoost. Logistic Regression had the best overall accuracy (0.96); however, it was incorrect in some instances, with some belonging to the minority class. Models that performed relatively well and best across the folds, with relatively little bias, were Random Forest and XGBoost, especially for Class 4. Overall, the Macro F1-Score indicates that the

model's performance decreases, leaving it imbalanced. The results are quite clear: Class 1, Class 2, and Class 3 are the most difficult to classify, with Class 4 doing slightly better but still not as good as Class 0.

4.2. Model Performance With SMOTE

The performance of the same models is presented in Table 4 after applying SMOTE to balance the dataset. The findings show improvements across all evaluation metrics.

Table 4. Performance of machine learning models after SMOTE.

Model	Class	Precision	Recall	F1-Score	Accuracy
Gaussian Naïve Bayes	0	0.75	0.93	0.83	0.82
	1	0.98	0.97	0.99	
	2	0.76	0.63	0.69	
	3	0.72	0.74	0.73	
	4	0.91	0.81	0.86	
Logistic Regression	0	0.92	0.90	0.91	0.88
	1	0.97	0.97	0.98	
	2	0.81	0.78	0.79	
	3	0.76	0.80	0.78	
	4	0.93	0.92	0.92	
Decision Tree	0	0.99	0.87	0.93	0.90
	1	0.98	0.98	0.97	
	2	0.92	0.78	0.84	
	3	0.70	0.96	0.81	
	4	0.98	0.88	0.93	
Random Forest	0	0.96	0.86	0.91	0.91
	1	0.97	0.98	0.99	
	2	0.89	0.86	0.88	
	3	0.80	0.96	0.87	
	4	0.97	0.89	0.93	
XGBoost	0	0.96	0.93	0.95	0.95
	1	0.98	0.98	0.99	
	2	0.87	0.88	0.87	
	3	0.89	0.91	0.90	
	4	0.97	0.94	0.96	

As shown in the results, the overall performance of evaluation metrics such as Precision, Recall, F1 score, and Accuracy has improved, indicating the effectiveness of SMOTE in reducing class imbalance and improving the learning ability of all models. While for the imbalanced dataset all models showed excellent performance, the performance of the models has weakened in Class 0. However, the models still performed well, with Precision ranging from 0.75 to 0.99, Recall from 0.86 to 0.93, and F1-score from 0.83 to 0.95; the dominance of the models in Class 0 has decreased compared to the imbalanced dataset, leading to more balanced decision boundaries. In Class 1, all models show very high Precision, Recall, and F1 score, almost 0.97 for most models and up to 0.99 for Gaussian Naïve Bayes, Random Forest, and XGBoost, which indicates that SMOTE is very effective in correctly labeling previously underrepresented samples. For Class 2, moderate to strong improvements are observed, with Decision Tree and Random Forest having the highest Precision (0.76 to 0.92) and Recall (0.63 to 0.88). At the same time, Gaussian Naïve Bayes also shows relatively lower Recall (0.63 to 0.88), which suggests that these classes are not as easy as Class 1. In the Class 3 category, Precision scores range from 0.70 to 0.89, Recall scores range from 0.74 to 0.96, and F1-score values range from 0.73 to 0.90, with Decision Tree and Random Forest having the highest Recall values of 0.96, indicating good detection capability. At the same time, XGBoost gives more balanced and stable scores. Finally, Class 4 has impressive and consistent performance across all models, with Precision of 0.91–0.98, Recall of 0.81–0.94, and F1-score of 0.86–0.96, with XGBoost and Decision Tree being the top-performing models, and some models performing at levels approaching that of Class 0, implying that Class 4 is now very distinct and well classified after

SMOTE. These results demonstrate overall the effectiveness of SMOTE to improve the classification performance on all classes and its ability to better balance Precision and Recall, with XGBoost and Random Forest giving the best and most stable performance. For a comparison, Table 5 summarizes the overall performance of the evaluated machine learning models before and after applying SMOTE.

Table 5. Overall comparison of the evaluated machine learning models before and after applying SMOTE

Model	Accuracy (Before SMOTE)	Accuracy (After SMOTE)
Gaussian Naïve Bayes	0.93	0.82
Logistic Regression	0.96	0.88
Decision Tree	0.90	0.90
Random Forest	0.94	0.91
XGBoost	0.95	0.95

As shown in Table 5, applying SMOTE improved classification performance for the minority classes across all evaluated models. Among the evaluated classifiers, XGBoost achieved the highest overall performance and provided more balanced predictions across all classes.

4.3.Confusion Matrix Analysis of SMOTE_ XGBoost Model

The confusion matrix analysis for the SMOTE_XGBoost model is presented in Fig. 3.

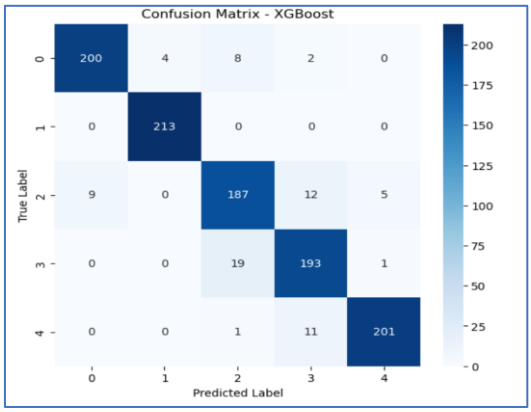

Figure 3: Confusion Matrix of XGBoost Model

The confusion matrix illustrates classification performance, with most predictions along the diagonal, indicating high accuracy for XGBoost. Class 1 is correctly classified, and Class 0 is also quite accurate, with a few misclassifications. Classes 2, 3, and 4 have good performance, with low confusion with neighboring classes, primarily because of the proximity of disease stages. In general, the findings indicate that XGBoost is a suitable machine learning algorithm for multiclass classification tasks, even when there is only a slight overlap between classes

4.4. Statistical Analysis

McNemar's test was used to assess the statistical significance of the XGBoost model's performance. This test quantifies the difference between two classifiers: b are cases classified correctly by XGBoost and incorrectly by the compared model, and c are cases the other way around, as defined in Table 6.

Table 6. Results of McNemar's Test - comparison of XGBoost with the other models

No	Model Comparison	b (XGB)	c (Other)	p-value
1	XGBoost vs Random Forest	53	33	3.99×10^{-2}
2	XGBoost vs Decision Tree	84	48	2.19×10^{-3}
3	XGBoost vs Logistic	96	41	2.95×10^{-6}
4	XGBoost vs Naïve Bayes	149	31	1.05×10^{-19}

The results show that all p-values are less than 0.05, indicating that the observed differences are significant. The performance of XGBoost is consistently better than Random Forest ($p = 0.039$), Decision Tree ($p = 0.002$), Logistic Regression ($p < 0.001$), and Naïve Bayes ($p \approx 0$), with the largest gap against Naïve Bayes.

5. DISCUSSION

Table 7 compares the proposed framework with representative studies on Hepatitis C prediction. Although previous studies reported competitive classification performance, direct comparison should be interpreted with caution because the studies differ in terms of datasets, preprocessing procedures, classification tasks, and evaluation protocols.

Table 7. Comparison of the proposed study with representative Hepatitis C prediction studies

Study	Dataset	Method	Accuracy
[11]	1801 clinical records	KNN + SHAP + SFS	83.1%
[12]	Public HCV dataset	Ensemble learning (MLP, Bayesian Network, QUEST)	95.59%
[13]	UCI Hepatitis dataset	Random Forest + Boruta	92.42%
[20]	Combined public HCV datasets	Meta-model with pre-clustering	94.82%
Proposed study	UCI HCV dataset	KNN imputation + Standard Scaler + SMOTE + XGBoost	95.00%, Macro F1 = 0.95

The experimental results indicate that class imbalance affected the performance of all evaluated machine learning models. Before applying SMOTE, the classifiers achieved high performance on the majority class, whereas lower Precision, Recall, and F1 scores were observed for the minority classes. After applying SMOTE, the classification performance of the minority classes improved across all evaluated models, indicating that data balancing reduced the effect of class imbalance and provided more representative decision boundaries. Out of the various classifiers that were tested, XGBoost had the best overall classification abilities as a result of adding SMOTE, while Random Forest had fairly consistent performance levels across the different disease classes. Previous studies have also indicated that ensemble-based models tend to be competitive for predicting Hepatitis C [12], [13], [20]. However, caution must be taken when comparing performance results due to differences between datasets, preprocessing methods, and experimental conditions. XGBoost was able to achieve such good prediction performance because it employs a gradient boosting technique, which creates multiple decision trees in order to minimize prediction error. This is done by fitting multiple decision trees in an iterative fashion to reduce the prediction errors from all of the trees combined. Gradient boosting can effectively capture complex nonlinear relationships among clinical variables while maintaining high generalization. Logistic Regression and Gaussian Naïve Bayes use simpler methods to describe the relationships between clinical laboratory tests in the dataset and therefore will likely not explain the many complex relationships within it. Before using SMOTE, the results indicate that the minority class of diseases was much more difficult to classify, due to the fact that there were very few training instances for these classes. Once the data was balanced by SMOTE, Precision, Recall, and F1-score were improved for most of the minority classes.

This proposed framework integrates KNN, imputation, feature standardization, SMOTE, and a comparative evaluation of both classical and ensemble machine learning models into a single preprocessing pipeline. This provides a unified experimental setting so that the influence of data preprocessing on multiclass Hepatitis C classification can be assessed through controlled comparisons of classifiers.

Study Limitations and Future Work. This study, however, has several limitations. First, the proposed framework was tested on a single publicly available dataset, and external validation was not performed. Although SMOTE positively influenced the classification of some minority classes, it may not be able to represent the distribution of real clinical data completely through synthetic

oversampling. Future work might include assessing the proposed framework on increasingly larger and more heterogeneous clinical datasets, incorporating explainable AI techniques, and exploring alternative data-balancing methods to assess the robustness and generalizability of the proposed approach.

6. CONCLUSION

In this research, we present a machine-learning framework for classifying multiclass Hepatitis C disease using appropriate clinical and laboratory data. Experimental findings demonstrate that data preprocessing, including class imbalance correction with SMOTE, enhances the classification performance of the evaluated machine learning models for minority classes. Results indicate that among the evaluated machine learning models, XGBoost achieves the best overall performance across all evaluations, while Random Forest also performs consistently across all assessed disease categories. In essence, appropriate preprocessing combined with ensemble learning has been established as a method to improve classification reliability for multiclass Hepatitis C disease. The proposed machine learning framework may also assist in clinical decision-making by distinguishing various stages of Hepatitis C disease. This framework is intended to support working with clinical evaluations and medical decision-making.

CONFLICT OF INTEREST

The authors declare that there is *no* conflict *of interest* regarding the publication of this paper.

REFERENCES

[1] J. Ferreira, J. Caldeira, M. Bicho, P. Faustino, and F. Serejo, "Hepatitis C Virus: An Overview of Its Chronic Impact on Liver Function, Metabolic Dysregulation, Inflammatory–Oxidative Pathogenesis and Epigenetic Memory," International Journal of Molecular Sciences, vol. 27, p. 3559, 2026, doi: 10.3390/ijms27083559.

[2] E. Pose, S. Piano, M. Thiele, N. Fabrellas, E. Tsochatzis, and P. Ginès, "Moving diagnosis of liver fibrosis into the community," Journal of Hepatology, vol. 83, 2025, doi: 10.1016/j.jhep.2025.01.026.

[3] H. Zilouchian, O. Faqah, M. Kabir, D. Gross, R. Pan, S. Shaifman, M. Younas, M. Haseeb, E. Thomas, and W. Asghar, "Current and Future Diagnostics for Hepatitis C Virus Infection," Chemosensors, vol. 13, p. 31, 2025, doi: 10.3390/chemosensors13020031.

[4] Y. Zhu, "Predicting hepatitis C infection via machine learning," American Journal of Translational Research, vol. 17, pp. 5120–5128, 2025, doi: 10.62347/QXZB5406.

[5] I. Shahid, A. Alzahrani, A. Al Ghamdi, I. Alanazi, S. Rehman, and S. Hassan, "Hepatitis C Diagnosis: Simplified Solutions, Predictive Barriers, and Future Promises," Diagnostics, vol. 11, p. 1253, 2021, doi: 10.3390/diagnostics11071253.

[6] D. H. H. Le, S. Kanokudom, H. Nguyen, R. Yorsaeng, S. Honsawek, S. Vongpunsawad, and Y. Poovorawan, "Hepatitis C Virus—Core Antigen: Implications in Diagnostic, Treatment Monitoring and Clinical Outcomes," Viruses, vol. 16, p. 1863, 2024, doi: 10.3390/v16121863.

[7] Y. Fahim, I. Hasani, S. Kabba, and W. Ragab, "Artificial intelligence in healthcare and medicine: clinical applications, therapeutic advances, and future perspectives," European Journal of Medical Research, vol. 30, 2025, doi: 10.1186/s40001-025-03196-w.

[8] S. Auger and G. Scott, "Machine learning detects hidden treatment response patterns only in the presence of comprehensive clinical phenotyping," PLOS One, vol. 20, 2025, doi: 10.1371/journal.pone.0334858.

[9] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, "Handling imbalanced medical datasets: review of a decade of research," Artificial Intelligence Review, vol. 57, 2024, doi: 10.1007/s10462-024-10884-2.

[10] A. Mohamed, M. Abdelrehim, and R. Al-Barazie, "Context matters in machine learning based disease prediction with insights from diverse clinical and symptom data," vol. 15, no. 1, p. 42669, 2025, doi: 10.1038/s41598-025-26855-8.

[11] A. Ali, M. Hassan, F. Aburub, M. Alauthman, A. Aldweesh, A. Al-Qerem, I. Jebreen, and A. Nabot, "Explainable Machine Learning Approach for Hepatitis C Diagnosis Using SFS Feature Selection," Machines, vol. 11, p. 391, 2023, doi: 10.3390/machines11030391.

[12] E. Onyema, S. Dalal, I. Ben Dhaou, C. C. Agubosim, C. Umoke, N. Richard-Nnabu, and N. Dahiya, "Artificial Intelligence-Based Ensemble Learning Model for Prediction of Hepatitis C Disease," Frontiers in Public Health, vol. 10, pp. 1–13, 2022, doi: 10.3389/fpubh.2022.892371.

[13] P. Khatun, S. Umam, R. B. Razzak et al., "A study on the effectiveness of machine learning models for hepatitis prediction," Scientific Reports, vol. 15, p. 30659, 2025, doi: 10.1038/s41598-025-07104-4.

[14] F. H. Yagin and A. Pinar, "Investigation of Hepatitis C diagnosis with machine learning and evaluation of clinical biomarkers with explainable artificial intelligence models," Medicine Science | International Medical Journal, vol. 14, p. 1309, 2025, doi: 10.5455/medscience.2025.04.092.

[15] A. Ahad, B. Das, M. R. Khan, N. Saha, A. Zahid, and M. Ahmad, "Multiclass Liver Disease Prediction with Adaptive Data Preprocessing and Ensemble Modeling," Results in Engineering, vol. 22, p. 102059, 2024, doi: 10.1016/j.rineng.2024.102059.

[16] T.-H. Li, H.-J. Chiu, and P.-H. Kuo, "Hepatitis C Virus Detection Model by Using Random Forest, Logistic-Regression and ABC Algorithm," IEEE Access, vol. 10, pp. 1–1, 2022, doi: 10.1109/ACCESS.2022.3202295.

[17] D. E. Nishwa, Z. Abbas, and S. W. Lee, "Machine learning-based prediction of treatment response in comorbid hepatitis C patients receiving DAA therapy: A real-world study from Pakistan," *Frontiers in Public Health*, vol. 14, Art. no. 1783217, 2026, doi: 10.3389/fpubh.2026.1783217.

[18] M. Hezari, M. Baes, A. Hezari, and M. Hassanbabaei, "Advanced Predictive Modeling for Hepatitis C Diagnosis Using Machine Learning," Clinical and Molecular Epidemiology, vol. 1, p. 12, 2024, doi: 10.53964/cme.2024012.

[19] S. A. Farooq, "The Multiclass Detection of Five Stages of Hepatitis C Using the Machine Learning Based Random Forest Algorithm," pp. 1–6, 2023, doi: 10.1109/WCONF58270.2023.10235157.

[20] A. Sharma, T. Khade, and S. M. Satapathy, "A cross dataset meta-model for hepatitis C detection using multi-dimensional pre-clustering," *Scientific Reports*, vol. 15, Art. no. 7278, 2025, doi: 10.1038/s41598-025-91298-0.

[21] S. K. Sunori *et al*., "Predictive Modelling of Hepatitis C Virus Disease Progression Using PCA and Machine Learning," *Biomedical & Pharmacology Journal*, vol. 19, no. 2, 2026.

[22] F. Soriano, "Hepatitis C Dataset," Kaggle, 2025. [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/hepatitis-c-dataset

[23] Lichtinghagen, F. Klawonn, and G. Hoffmann. "HCV data," UCI Machine Learning Repository, 2020. [Online]. Available: https://doi.org/10.24432/C5D612.

[24] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," Scientific Reports, vol. 14, 2024, doi: 10.1038/s41598-024-56706-x.

[25] C. Miller, T. Portlock, D. Nyaga, and J. O'Sullivan, "A review of model evaluation metrics for machine learning in genetics and genomics," Frontiers in Bioinformatics, vol. 4, 2024, doi: 10.3389/fbinf.2024.1457619.