I.J. Information Technology and Computer Science, **2026, 2, 161-182**
Published Online on April 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2026.02.10

Hepatitis C Diagnosis using Supervised Machine Learning Algorithms and Ensemble Learning Techniques

Karthika Natarajan*
School of Computer Science and Engineering, VIT-AP University, Amaravathi, 522237, India
E-mail: karthika.n@vitap.ac.in
ORCID iD: https://orcid.org/0000-0002-6700-583X
*Corresponding author

Koteswara Rao Makkena
School of Computer Science and Engineering, VIT-AP University, Amaravathi, 522237, India
E-mail: makkena.21phd7131@vitap.ac.in
ORCID iD: https://orcid.org/0000-0001-6748-3420

Received: 03 November 2025; Revised: 28 January 2026; Accepted: 24 February 2026; Published: 08 April 2026

Abstract: Hepatitis, a severe and highly impactful disease, poses significant challenges for healthcare systems, including limited diagnostic resources, delayed detection, and inadequate treatment infrastructure. This work addresses these issues by developing a machine-learning predictive system to classify hepatitis severity. By employing Logistic Regression, Random Forest, SVM, KNN, and ensemble techniques such as AdaBoost, CatBoost, and Gradient Boosting, the system enhances early detection and severity assessment. The issue of class imbalance was addressed using ADASYN and SMOTE methods applied to two separate datasets. For Dataset 1, following the use of the ADASYN technique, the achieved accuracies were 88.11% for Logistic Regression, 98.92% for Random Forest, 97.30% for AdaBoost, and 96.22% for Gradient Boosting. When SMOTE was employed on Dataset 1, Random Forest and Gradient Boosting reached accuracies of 98.38% and 96.76%, respectively. In the case of Dataset 2, AdaBoost achieved an accuracy of 93.75% after applying both ADASYN and SMOTE. These models analyze clinical data to deliver accurate, timely predictions, reducing the burden on resource-constrained healthcare systems. Ensemble methods enhance model robustness and accuracy, supporting improved decision-making and efficient resource allocation. Furthermore, SHAP offers global explanations of feature importance and force plots for local interpretations, while LIME increases the interpretability of results from black-box models, facilitating effective hepatitis management. Future work will focus on integrating interoperability standards, such as HL7 FHIR, to enable real-time data exchange, facilitating seamless risk assessment and clinical decision support within healthcare workflows.

Index Terms: Hepatitis C, SMOTE, ADASYN, Random Forest Classifier, Adaboost Classifier, SHAP, LIME.

1. Introduction

Liver disease is globally recognized as a significant chronic ailment affecting the liver. Among the human body's organs, the liver stands out as one of the largest and heaviest [1]. Its functions include bile secretion, protein synthesis, and detoxification. Consequently, inflammation of the liver due to hepatitis results in compromised liver function, leading to a decline in the individual's overall health. Hepatitis symptoms vary among patients, and some individuals may exhibit no discernible signs [2]. Typical manifestations comprise jaundiced eyes and skin, abdominal distress, diminished appetite, and fatigue. Hepatitis C is a virus that spreads through contact with infected blood, with most infections stemming from blood exposure due to hazardous injection instruments, inadequate healthcare procedures, blood transfusions, intravenous drug administration, and sexual behaviors associated with blood contact [3]. It can be categorized as short-term or long-term depending on its duration. Conditions lasting less than six months are classified as short-term, whereas those persisting for more than six months are considered long-term. [4]. Conventional methods for diagnosing hepatitis can be challenging and often entail costly medical tests [5]. Furthermore, diagnosing such diseases using intelligent systems not only lowers costs but also expedites the patient examination process. Therefore, the creation of intelligent diagnostic systems for predicting such diseases is of paramount importance. Table 1 below

This work is open access and licensed under the Creative Commons CC BY 4.0 License.

illustrates the classification of various hepatitis types into acute and chronic categories [6].

All five types of hepatitis must be treated since they cause infections and many deaths. Notably, hepatitis types B and C are the primary contributors to liver cirrhosis, cancer, and chronic illnesses affecting millions of individuals worldwide [7]. As reported by the global health agency World Health Organization (WHO), approximately 130 to 150 million individuals across the globe are affected by long-term hepatitis C infection [8]. This disease can be transmitted using contaminated tools like syringes and cutting implements, contributing to the dissemination of the infection.

While each of the five hepatitis virus variants has the potential to cause severe diseases, the highest number of global fatalities can be attributed to chronic infections of hepatitis B or hepatitis C [9]. The driving force behind this endeavour is to minimize human errors in the diagnostic process.

Although the proposed model shows strong predictive performance, its effective deployment in real-world clinical environments, especially in under-resourced regions, demands careful attention to integration and infrastructure challenges. Classification models are utilized for disease classification. Machine Learning (ML) Classification models like Logistic Regression (LR)[10], Random Forest (RF) [11], Support vector Machine (SVM) [12], K-Nearest Neighbors (KNN) [13] and EL(EL) Techniques like AdaBoost [14], CatBoost [15], Gradient boosting [16]. The method for discovering hepatitis disease described in this paper integrates ML and EL techniques.

To evaluate the method, the authors employed the Kaggle hepatitis C datasets [17–18]. The method has been introduced with the aim of elevating the diagnostic accuracy beyond that of existing methods. The subsequent are the major contributions of the proposed method.

- Through the application of Adaptive Synthetic Sampling (ADASYN) technique to rectify unbalance of data in a dataset1 and dataset 2, conducted a comparative analysis of diverse Evaluation metrics across various ML classification models and EL techniques.
- The proposed a method for diagnosing hepatitis C disease, which involves identifying the most efficient classifier among ML classification models and EL models after implementing Synthetic Minority Over-sampling Technique (SMOTE) balancing technique.
- SHapley Additive exPlanations (SHAP) implementation provides global explanations for feature importance and force plots for local explanations, enhancing the interpretability of results from black-box models for effective hepatitis management.

Manuscript Structure

The structure of this manuscript is as follows: Section 2 provides a comprehensive literature review. Section 3 describes the proposed method for hepatitis c disease. Section 4 outlines procedures and materials employed in this work. Section 5 presents ML models utilized, while Section 6 discusses the EL techniques implemented. Section 7 addresses class balancing techniques. Section 8 performance and evaluation, and Section 9 presents the results and discussion. Finally, Section 10 conclusion and future scope.

2. Literature Review

Based on their findings, researchers conducted in several of studies that assist to diagnose hepatitis. Xiaolu Tian and colleagues [19] have pinpointed the most effective model for forecasting hepatitis B surface antigen serological resolution. Among the evaluated models, XGBoost exhibited the highest predictive performance, achieving an AUC of 0.891, followed by RF (0.829), DT (0.619), and LR (0.680).

N Komal Kumar et al [20] analyzed and predicting the infectious hepatitis disease by comparing the performance of the classifiers LR, RF, DT, C4.5 and MLP. The evaluation metric is derived from True Positive Rate (TPR) and accuracy. At an accuracy of 90.32%, the RF model attained the highest performance compared in the forecasting of infectious hepatitis.

Vedha Krishna Yarasuri et al. [21] designed a tool for diagnosing and detecting hepatitis, along with predicting the life expectancy of diseased patients. The study involved comparing different machine learning models and neural networks, specifically SVM, KNN, and ANN, as a tools for diagnosing disease. Achieving the lowest mean squared error and a high prediction accuracy of 96%, the ANN model emerged as the most accurate among all the evaluated algorithms.

Azam Orooji et al. [22] developed a prediction system for HCV disease and categorized its severity using ML models. Three ML models MLP, NB, and DT were employed, and a cross-validation with 10-fold method was utilized to assess machine learning models. The utilization of an oversampling technique demonstrated that RF got the highest accuracy when compared to other models.

Matko Glucina et al [23] proposed a system for diagnosing Cervical cancer. Class balancing techniques SMOTE, ADASYN, SMOTEEN, random oversampling, and SMOTE Tomek were used for balancing dataset. In this study ML classifiers such as KNN, LR, MLP, SVM and NB methods were implemented. The findings reveal enhanced performance with the combination of MLP and KNN, using Random Sampling, SMOTEEN, and SMOTE Tomek techniques. This approach consistently yields AUC and MCC mean scores exceeding 0.95, irrespective of the chosen diagnostic method for output vector construction.

Ravi Kumar Sachdeva et al [24] devised a systematic model for diagnosing hepatitis disease employing ML classifiers such as KNN, RF, LR, and SVM. The evaluation of these classifiers was conducted both with class imbalance and with the incorporation of SMOTE strategy for class balance. Out of these methods, LR with SMOTE showcased the highest accuracy, reaching 93.18%.

Ayşe Doğru and others [25] showed a new super-EL model that can help find diabetes mellitus early. It uses SVM, DT, RF, gradient boosting, LR. The new model was more accurate than individual base learners. It did especially well in predicting early-stage diabetes risk, the PIMA dataset, and the diabetes 130-US hospitals dataset.

Nahid Hossain Taz et al. [26] developed a diabetes prediction framework using ensemble-based ML on the PIDD. A study of the performance of different classifiers (LightGBM, XGBoost, AdaBoost, and RF) showed that LightGBM had the best accuracy 94% and ROC AUC score of 95%. An SV classifier made the overall accuracy structure even better 95% and ROC AUC score of 96%.

Zahra Sadeghi et al. [27] explored the elements and challenges of XAI in health care through a comprehensive review. The study aims to review diverse XAI methods, their associated challenges, and ML models in healthcare, highlighting XAI's significant contribution to solving healthcare problems, especially in safety-critical applications

Ali Mohd Ali et al. [28] aimed to evaluate the effectiveness of multiple ML classifiers in diagnosing liver disease, particularly hepatitis C, with the goal of improving diagnostic speed and minimizing associated costs. The study utilized the Sequential Forward Selection (SFS) method to enhance model performance by identify the most influential features. Results demonstrated that, across all classifiers tested on the dataset, the average accuracy reached approximately 83%.

The following table differentiation of existing work could be better emphasized to highlight its unique contribution of proposed system.

Table 1. Comprehensive performance comparison of existing machine learning models and the proposed system based on classification accuracy

Sno	Author	Accuracy for ML Models
1	Xiaolu Tian et al[18]	XGBoost-89.10, RF-82.90, DT-61.90, and LR-68.0
2	N Komal Kumar et al[19]	LR-83.8, RF-90.32, DT-87.09, C4.5 -83.87 and MLP-75.80
3	Vedha Krishna Yarasuri et al[20]	SVM-89.5, KNN-85.29, and ANN-96.153
4	Azam Orooji et al[21]	MLP-96.40, RF-90.00, and J48-98.01
5	Ravi Kumar Sachdeva et al[23]	KNN-81.25, RF-85.00, LR, -85.00 and SVM-83.7
6	**Proposed System**	LR-90.27, **RF-98.38**, SVM-94.5, KNN-89.73, AdaBoost-96.76, CatBoost- 95.14 Gradient Boost-96.76%

3. Proposed Method for Hepatitis C Disease

The proposed method is illustrated in Figure 1. An imbalanced dataset was subjected to various classification algorithms and EL techniques. The author applied ADASYN & SMOTE for class balance. After class balancing using both ADASYN and SMOTE, classification techniques like LR, RF, SVM, KNN and EL techniques were used to diagnosis hepatitis C disease.

After applying the median to handle missing values in a dataset, it revealed 384 occurrences of 'not affected' and 46 occurrences of 'affected,' indicating an imbalanced dataset1. The author applied ADASYN for class balancing 384 occurrences of not effected and 380 occurrences are affected. The author applied SMOTE for class balancing 384 occurrences of not effected and 384 occurrences are affected. Here is the pseudocode utilized for implementing the method depicted in Figure 2.

In proposed methodology, as depicted in Figure 1, employs a modular and scalable framework for liver disease severity classification. The pipeline integrates efficient preprocessing, feature selection, and class balancing techniques, followed by classification using EL methods. This work addresses class imbalance by applying ADASYN and SMOTE, leading to improved model robustness across diverse data distributions.

For real-time clinical deployment, this study focuses on enhancing model performance and efficiency through the following optimization strategies:

- Model pruning and quantization: Reduces memory and computational overhead, improving inference speed and operational efficiency.
- Lightweight ensemble models: Utilizes shallow decision trees or gradient boosting with constrained estimators to ensure low resource consumption.
- Deployment compatibility: Integrates models as RESTful APIs or embedded modules compatible with existing Electronic Health Record (EHR) systems and point-of-care devices.

The preprocessing and feature selection processes, once established, require only a limited set of inputs—such as liver enzyme levels and standard laboratory parameters, allowing for efficient real-time severity classification. Although the current study concentrates on model development and validation, seamless integration with Electronic Health Record (EHR) systems remains crucial for real-world implementation. Future work will focus on achieving interoperability

through HL7 FHIR standards to facilitate real-time data exchange, enabling dynamic risk assessment and clinical decision support within existing healthcare workflows.

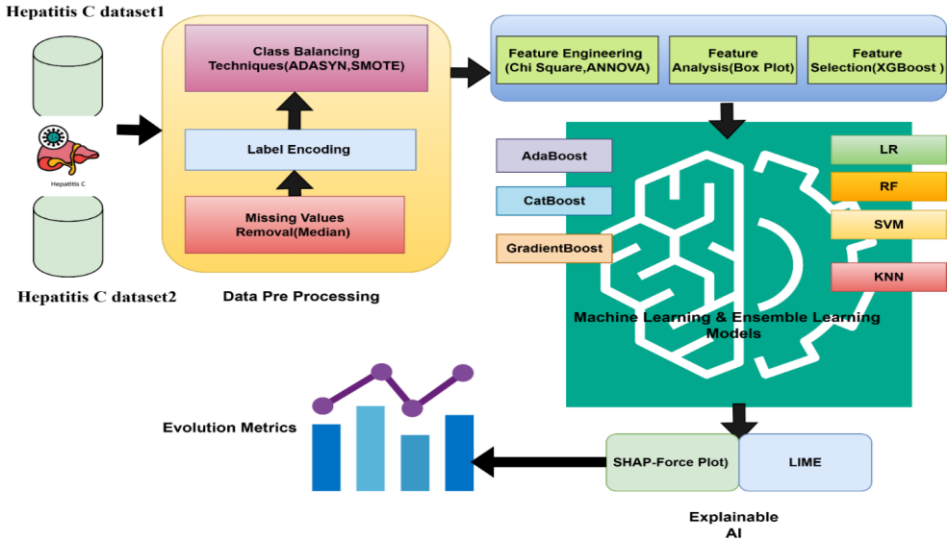


Fig.1. Proposed method for hepatitis disease diagnosis

4. Procedures and Materials

4.1. Data Set

This study utilized two publicly available Hepatitis C datasets obtained from Kaggle [17–18]. The first dataset (Dataset 1) includes 615 patient records with 14 features, while the second dataset (Dataset 2) comprises 155 records with 21 features. Both datasets encompass demographic factors (such as age and sex) along with clinical and laboratory parameters. Dataset 1 categorizes subjects into four groups, Blood Donor, Hepatitis, Fibrosis, and Cirrhosis—thereby enabling stratified analysis across different stages of liver disease progression. Although hepatitis subtypes (A, B, C) are not explicitly identified, Dataset 1 remains highly relevant for generalized liver disease prediction. The target variable in Dataset 1, labeled 'Category', is binary, with 1 indicating disease presence and 0 indicating absence. Similarly, in Dataset 2, the target variable 'target' follows the same binary classification scheme. Tables 2 and 3 provide detailed descriptions of the features included in Hepatitis C Dataset 1 and Dataset 2, respectively.

Table 2. Detailed description of features in hepatitis c dataset-1 including clinical attributes, value ranges, and data type discretization

Sno	Feature	Name	Values	Discretization
1	Age	Age of the patient	~18 to 80	Continuous Numerical
2	Sex	Biological sex	m, f	Categorical
3	ALB	Albumin	20.0 to 55.0 g/L	Continuous Numerical
4	ALP	Alkaline Phosphatase	30.0 to 300.0 U/L	Continuous Numerical
5	ALT	Alanine Aminotransferase	10.0 to 400.0 U/L	Continuous Numerical
6	AST	Aspartate Aminotransferase	10.0 to 300.0 U/L	Continuous Numerical
7	BIL	Bilirubin	0.1 to 5.0 mg/dL	Continuous Numerical
8	CHE	Cholinesterase	3.0 to 13.0 kU/L	Continuous Numerical
9	CHOL	Cholesterol	100.0 to 300.0 mg/dL	Continuous Numerical
10	CREA	Creatinine	0.5 to 1.8 mg/dL	Continuous Numerical
11	GGT	Gamma-Glutamyl Transferase	10.0 to 200.0 U/L	Continuous Numerical
12	PROT	Total Protein	5.0 to 9.0 g/dL	Continuous Numerical
13	Category	Diagnostic Class	0: Not Effected, 1: Effected	Categorical (Binarized)

Table 3. Detailed description of features in hepatitis c dataset-2 including patient symptoms, laboratory parameters, and data type discretization

Sno	Feature Name	Values	Discretization
1	age	Age of the patient (~7 to 78)	Continuous Numerical
2	gender	Biological sex (1: Male, 2: Female)	Categorical
3	steroid	Use of steroids (1: Yes, 2: No)	Categorical
4	antivirals	Use of antivirals (1: Yes, 2: No)	Categorical
5	fatigue	Presence of fatigue (1: Yes, 2: No)	Categorical
6	malaise	Presence of malaise (1: Yes, 2: No)	Categorical
7	anorexia	Presence of anorexia (1: Yes, 2: No)	Categorical
8	liverBig	Enlarged liver (1: Yes, 2: No)	Categorical
9	liverFirm	Firm liver (1: Yes, 2: No)	Categorical
10	spleen	Enlarged spleen (1: Yes, 2: No)	Categorical
11	spiders	Presence of spider nevi (1: Yes, 2: No)	Categorical
12	ascites	Presence of ascites (1: Yes, 2: No)	Categorical
13	varices	Presence of varices (1: Yes, 2: No)	Categorical
14	bili	Bilirubin (e.g., 0.3 to 8.0 mg/dL)	Continuous Numerical
15	alk	Alkaline Phosphatase (e.g., 30–300 U/L)	Continuous Numerical
16	sgot	SGOT (AST) (e.g., 10–300 U/L)	Continuous Numerical
17	albu	Albumin (e.g., 2.0 to 5.5 g/dL)	Continuous Numerical
18	protime	Prothrombin time (e.g., 0–100 seconds)	Continuous Numerical
19	histology	Histological liver damage (1: Yes, 2: No)	Categorical
20	target	Diagnosis (0: Not effected, 1: effected)	Categorical (Binarized)

4.2. Development Environments

A. Google Colab

The cloud-based Python notebook environment is used for ML model development and testing, with free access to GPUs and TPUs for heavy computations and facilitated real-time collaboration.

B. Local System Environment

The local system is equipped with an AMD Ryzen 7-5800H with Radeon Graphics CPU, which operates at a 3.2 GHz base frequency ensuring smooth multitasking. Installed in the system are 16 GB of RAM (9.2 GB usable), found adequate for handling medium-sized datasets and allowing for the delay-free running of ML algorithms. The device is powered by a 64-bit operating system and utilizing Windows 11 Home Single Language, version 10.0, furnished with an OS Build 22631 and a Windows Feature Experience Pack to enhance user experience and security.

4.3. Data Preprocessing

In Dataset 1, the original Category attribute was binarized to differentiate healthy blood donors (0) from patients with liver conditions, including Hepatitis, Fibrosis, and Cirrhosis (1). Non-informative columns, such as Unnamed: 0, were removed prior to analysis. Missing values were detected through null value inspection and imputed using the median to preserve the dataset's distribution and reduce the impact of outliers. Categorical variables, such as Sex, were label-encoded ('m' = 1, 'f' = 2) to ensure compatibility with machine learning algorithms. Feature types were further classified based on the number of unique values, where attributes with more than six unique entries were treated as numerical, and those with fewer were considered categorical.

In Dataset 2, the target attribute was binarized to differentiate healthy blood donors (0) from patients diagnosed with liver conditions such as Hepatitis (1). Non-informative columns, including ID, were removed to eliminate redundancy. Categorical variables such as Sex were label-encoded ('f' = 1, 'm' = 2) to ensure compatibility with machine learning algorithms. Features were classified based on the number of unique values, where those with six or more unique entries were treated as numerical, while the remaining ones were considered categorical. Chi-squared tests were applied to identify statistically significant categorical features, whereas ANOVA was employed for numerical feature selection. Additionally, boxplot analysis was conducted to visualize feature distributions and detect potential outliers. The implemented preprocessing pipeline ensures clean, normalized, and representative input data, enhancing both the robustness and interpretability of the machine learning models. Figure 2 illustrates the distribution of suspected patients and healthy individuals, with Dataset 1 comprising 540 suspected cases and 75 classified healthy controls.

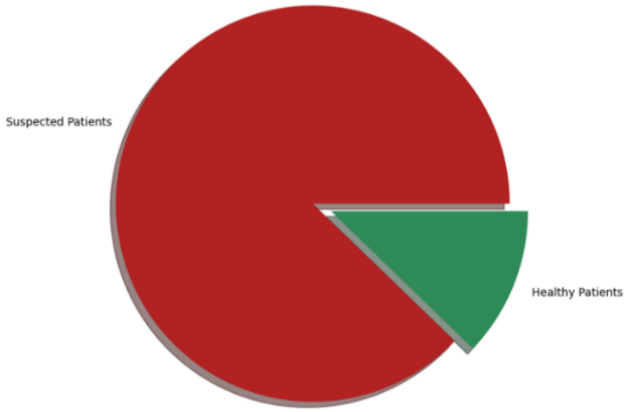


Fig.2. Plot for Suspected Patients vs Healthy Patients on dataset 1

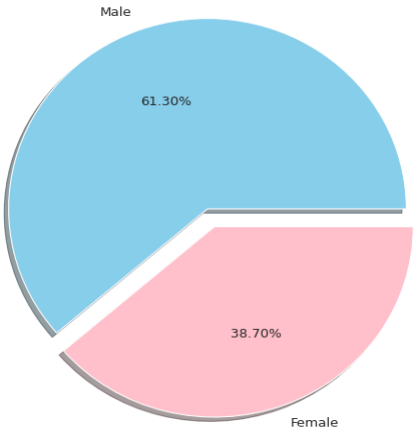


Fig.3. Patient data categorized by gender on dataset 1

Figure 3 presents the distribution of patients by gender within the dataset 1. Here '1' indicates for male and '2' indicates for female in this preprocessing stage. In the dataset 1 male patients: 377(61.30%) and female patients: 238(38.70%).

The following Figure.4 shows suspected patients and healthy patients. The dataset 2 comprises 123 individuals identified as potential cases and 32 individuals classified as healthy controls. Figure 5 presents the distribution of patients by gender within the dataset 2. Here '2' indicates for male and '1 indicates for female in this preprocessing stage. In the dataset 2 male patients :89.68% and female patients: 10.32%.

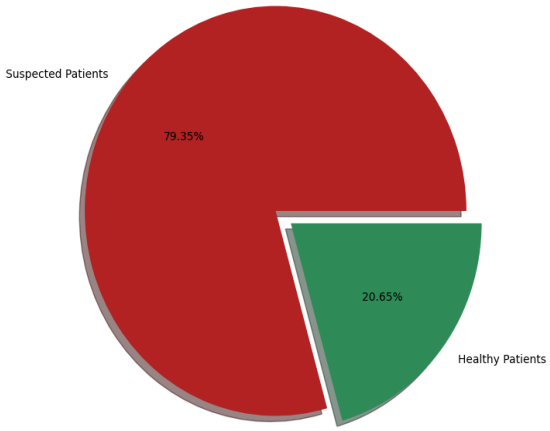


Fig.4. Plot for suspected patients vs healthy patients on dataset 2

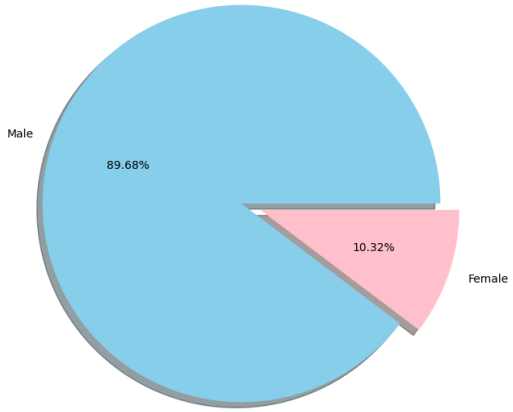


Fig.5. Patient data categorized by gender on dataset 2

To acknowledge the importance of quantifying class imbalance, particularly in medical datasets where the minority class (e.g., affected/severe cases) holds greater clinical significance.

In original dataset 1 out of a total of 615 patients, 540 were unaffected and 75 were affected, resulting in a class distribution of approximately 87.8% unaffected and 12.2% affected. Class imbalance can result in inflated accuracy when a model defaults for detecting the majority category. To address the problem of class imbalance, SMOTE and ADASYN techniques were applied. To ensure reliable assessment in imbalanced scenarios, the model was evaluated using evaluation metrics.

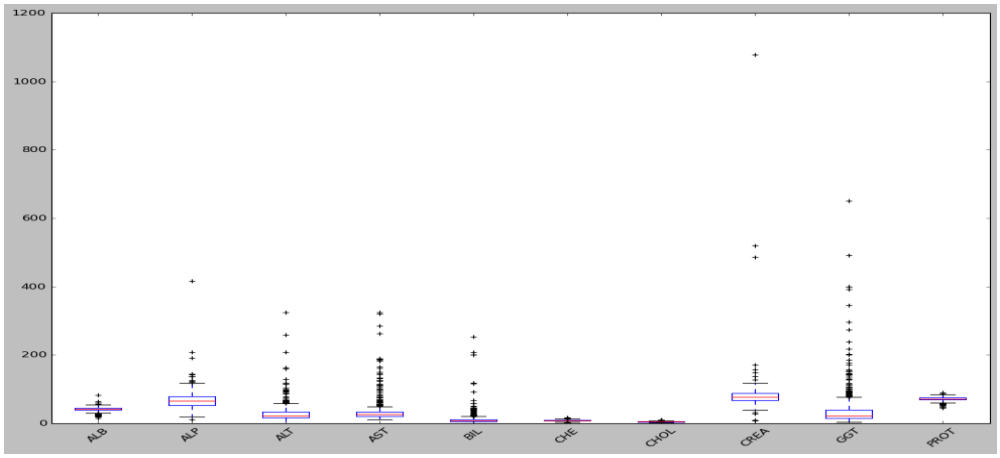


Fig.6. Boxplot for independent variables of Dataset 1

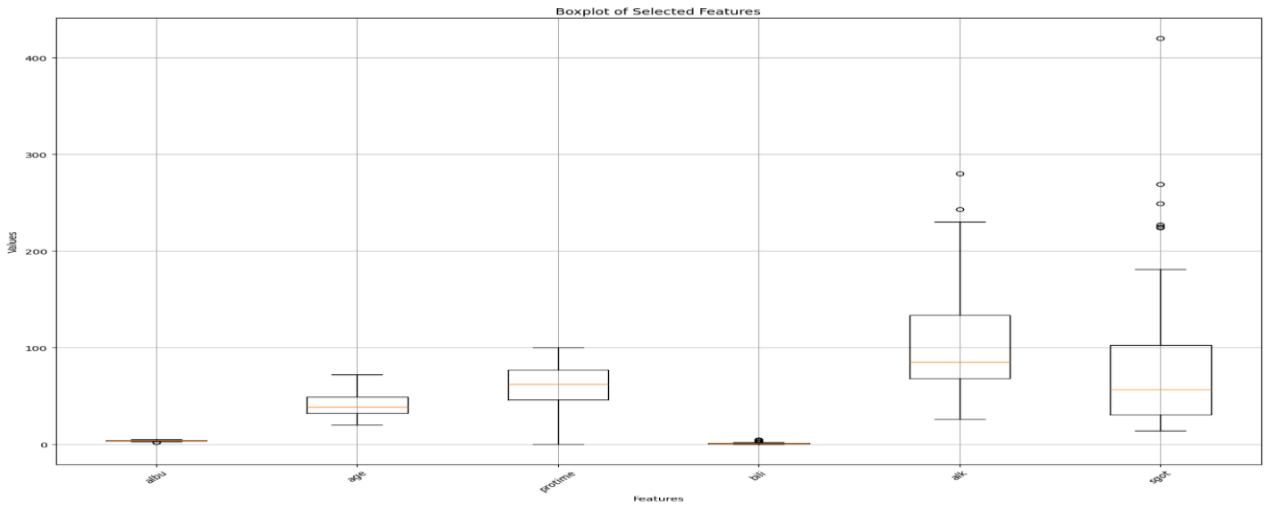


Fig.7. Boxplot for independent variables of Dataset 2

4.4. Feature Analysis

The boxplot analysis provides a visual summary related to the distribution and diversity of numeric characteristics [30]. Each plot illustrates in Figure 6 the median, interquartile range, and the presence of outliers for clinical attributes. Notably features like ALT, AST, and GGT exhibit significant skewness with a high number of outliers, indicating substantial variation in liver enzyme levels, which may correspond to disease severity.

Based on Figure 7, each boxplot illustrates the median, interquartile range, and presence of outliers for selected clinical attributes. Features such as alk and sgot show considerable skewness and many outliers, indicating significant variability in liver enzyme levels. This variation may reflect differing degrees of liver dysfunction or disease severity among the patients.

4.5. Feature Selection Techniques

Multiple feature importance techniques, such as chi-squared, ANOVA, XGBoost, and RF, were applied to determine the most impactful features. Features were selected using these feature importance methods.

A. Anova

To identify the most informative numerical features, ANOVA (Analysis of Variance) was applied to evaluate their statistical relevance with respect to the class label [31]. As shown in the Figure 8, features such as CHE (18.05), ALT (16.42), and CREA (16.02) achieved the highest ANOVA scores, indicating strong discriminatory power between healthy individuals and patients with liver conditions in dataset 1. Features like ALB, GGT, and AST also showed moderate influence, justifying their inclusion in the model. Conversely, variables such as CHOL, Age, and ALP recorded very low ANOVA scores, suggesting limited contribution to class separation. These results guided the selection of features most predictive of disease status, thereby improving model focus and reducing overfitting.

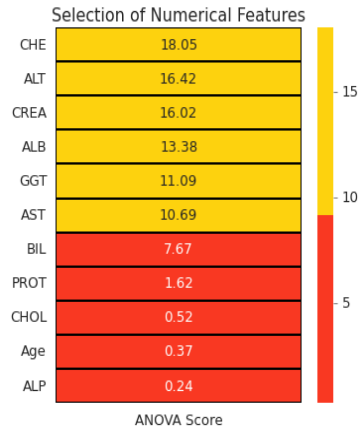


Fig.8.ANOVA score on dataset 1

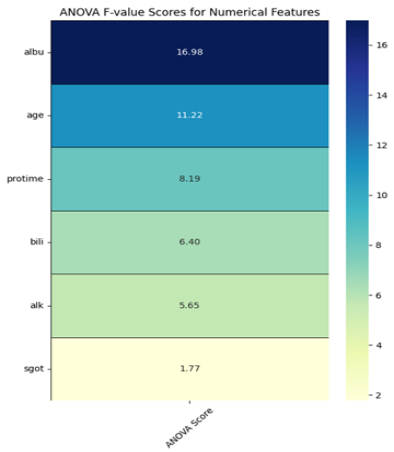


Fig.9. ANOVA score on dataset 2

As shown in the below Figure 9 heatmap, albu (16.98), age (11.22), and protime (8.19) achieved the highest ANOVA F-values, indicating strong discriminatory power among the numerical features. These attributes are likely to contribute significantly to differentiating between clinical outcomes. Features such as bili (6.40) and alk (5.65) showed moderate influence, supporting their inclusion in the model. In contrast, sgot recorded the lowest ANOVA score (1.77), suggesting

limited relevance for class separation. These results helped prioritize the most informative features, enhancing predictive accuracy and minimizing model complexity.

B. Chi-square

The next section elaborates on the operational concepts of the chi-square approach since features produced by this method yield the greatest accuracy score [32].

The chi-square test assesses the independence between variables and is commonly employed for feature selection in classification scenarios with categorical target variables. The below equation represents chi-square.

$$X^2 = \sum \frac{(O-E)^2}{E} \tag{1}$$

O indicates the observed value; E corresponds to the expected value

C. XGBoost

XGBoost, an optimized gradient boosting framework is frequently used in supervised learning due to its strong predictive performance [33].

$$\text{DS}=\{(x_i, y_i): i = 1 \dots \dots . n, x_i \in \mathbb{R}^m, y_i \in \mathbb{R}\} \tag{2}$$

DS is the dataset that has m features, then for n instances,

Let yˆi be the output value of the boosting model, specified using below equation.

$$y^i = \emptyset(x_i) = \sum_{k=1}^{K} f_{k\,(x_i)}, f_k \in \mathcal{F} \tag{3}$$

In this model, K indicates the number of decision trees, with f_k referring to the k-th tree. Feature importance for the Hepatitis C dataset is assessed using the XGBoost classifier.

XGBoost classifier addressed the interpretability challenge by integrating SHAP to provide local (instance-level) explanations for model predictions. It provides visual tools like force plots that clearly display how different features increase or decrease the prediction value. This makes the model's decisions easier to interpret, allowing clinicians to understand the rationale behind a particular result, thereby fostering greater transparency and acceptance of the system.

D. SHAP (SHapley Additive exPlanations)

To analyze results of a ML model, SHAP uses a game theory method. Using the Shapley values principle of cooperative game theory, it assesses the relative significance of numerous factors that determine the prediction. The Shapley plot is a useful tool that SHAP offers for showing Shapley values. For each dataset data point, these graphic representations illustrate how every feature contributes to the model's prediction. Classifying the most essential features in the upper area if the plot relies on their significance. As a model-agnostic tool, SHAP enables explanation of predictions across various black-box models; this is one of its key benefits. In addition, SHAP simply explains the model's predictions in specific instances, or "local explanations," rather than the whole thing [34].

SHAP leverages Shapley values to quantify the contribution of each feature in complex machine learning models. The concept of Shapley values entails computing the (aggregate) average of individual contributions. The definition is determined by the influence of the value of a feature on the prediction across all potential groupings or coalitions.

SHAP is specified as

$$\varphi i(x) = \sum_{s \subseteq \{1,\dots,p\}\{i\}} \frac{|S|!(p-|S|-1)!}{p!}[val(S \cup \{I\}) - val(S)] \tag{4}$$

$$\varphi i(x) = s \subseteq \{1,\dots,p\}\{i\} \quad S!p-S-1!p![valS \cup I-\text{val(S)}] \tag{4}$$

Here, $\varphi_i(x)$ denotes the Shapley value assigned to feature *i*; *I* represent the feature value, and *S* signifies a subset (coalition) of features. *p represents the count of model features, and val denotes the prediction for the feature value within set S.*The following formula specifies the value of a function for a particular instance x.

$$val_x(S) = \int f(x_1, x_2, \dots, x_p)\, dP_{x \notin S} \tag{5}$$

Value function indicates model prediction marginalized over feature values not in the coalition.

a. Global Explanations

Fig.10 demonstrates a SHAP-based summary visualization that highlights the overall impact of features in the

XGBoost model for predicting Hepatitis C. It evaluates the importance of input variables in shaping methods predictions and visually demonstrates the effect of their value changes Liver enzymes such as AST, GGT, and ALT emerged as the most influential predictors, with higher values significantly increasing the likelihood of a positive Hepatitis C diagnosis. Likewise, elevated levels of bilirubin and alkaline phosphatase were positively associated with the model's output. In contrast, lower values of these biomarkers tended to push the predictions toward the non-infected class. The color gradient in the plot represents the actual magnitude of each feature, providing an intuitive view of how specific values influence model decisions. This SHAP-based interpretation enhances the transparency of the XGBoost classifier and supports its reliability in clinical decision-making for Hepatitis C detection.

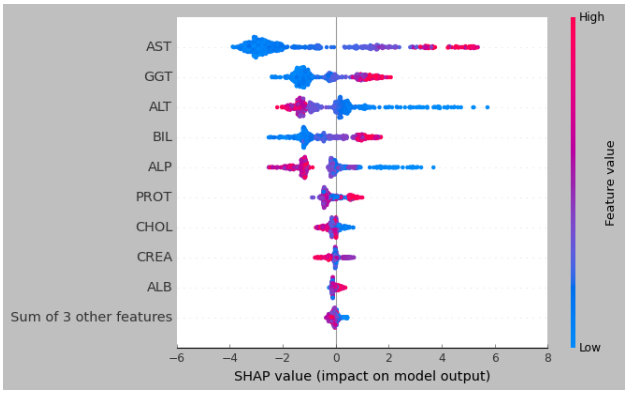


Fig.10. Global feature importance for hepatitis c prediction

b. Local Explanations

The SHAP force plot presented in Figure 9 provides a detailed interpretation of the Hepatitis C prediction for an individual patient by illustrating how each clinical feature influences the model's output. The base value denotes the mean model prediction across all patients, while the specific prediction for this instance (f(x) = –6.75) indicates a strong likelihood of being non-infected as shown in Figure 11. Features such as total protein (PROT), cholesterol (CHOL), aspartate transaminase (AST), bilirubin (BIL), gamma-glutamyl transferase (GGT), and alkaline phosphatase (ALP) exerted negative contributions, driving the prediction toward the non-infected category. In contrast, age and alanine transaminase (ALT) showed positive contributions, suggesting a mild risk, though insufficient to offset the dominant negative effects of other variables. This analysis demonstrates how SHAP values provide transparent and clinically meaningful insights into the decision-making process of the Hepatitis C predictive model.

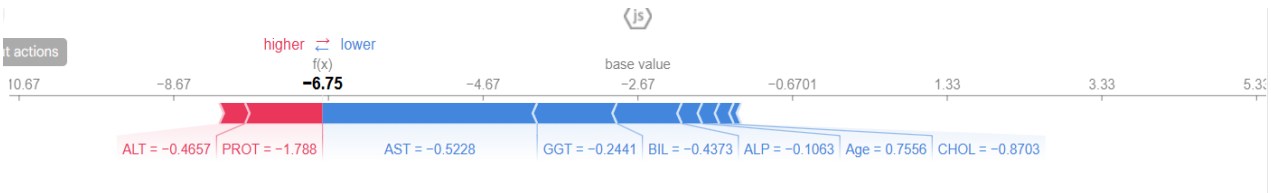


Fig.11. Local feature contributions for a single hepatitis c prediction

This study enhanced the model's clinical reliability by integrating Explainable AI techniques. SHAP force plots were employed to interpret individual predictions, allowing clinicians to visualize how factors such as bilirubin levels, enzyme concentrations, and patient age influenced the final decision. This approach promotes trust and transparency in the model's predictions while ensuring that performance on minority classes remains both clinically meaningful and interpretable, particularly in high-stakes medical applications.

c. LIME

A local replacement model, rather of a global one, can be trained with the help of LIME to explain the individual outcome inference for cirrhosis patents conducted by the complex ML models [35].

The model confidently predicted that the patient belongs to Class 0 (Effected) with a probability of 99%, indicating a strong likelihood of being affected by the condition. As illustrated in Figure 12 the LIME explanation highlights the key features contributing to this prediction. Among the most influential features was the patient's age (48 years), which falls within the range 39.5 to 50, a range strongly associated with the "Effected" class in the model. Additionally, the gender (value = 1) and the use of antivirals (value = 2) also contributed significantly toward the same class. Other features such as the presence of fatigue, malaise, and anorexia, along with conditions like enlarged liver (liverBig = 2), firm liver (liverFirm = 1), and spleen enlargement (spleen = 2), provided smaller yet supportive contributions toward the affected class prediction. Notably, none of the features showed influence toward Class 1 (Not Effected), which further reinforces the model's high confidence in classifying the patient as affected. Overall, this localized interpretation offers a clear

understanding of the influence that individual feature values had on the prediction for this case. LIME explanation for more instances also from the hepatitis C dataset 1.

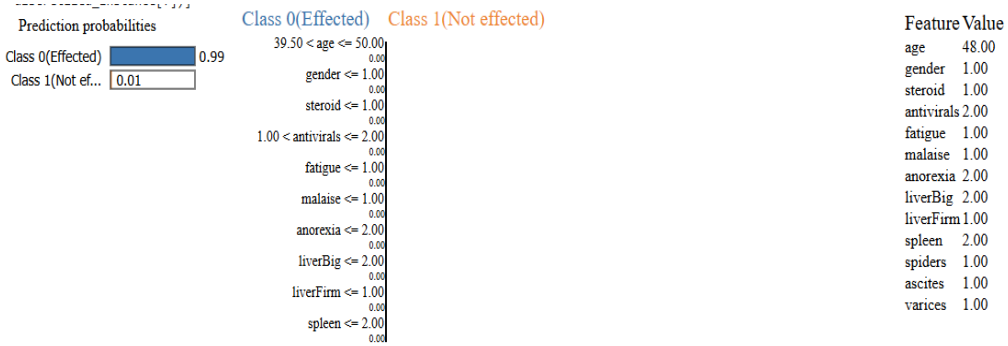


Fig.12. LIME explanation for first instance from given hepatitis C dataset 2

5. Machine Learning (ML) Models

The method under consideration employs four distinct classifiers for the purpose of easy diagnosis hepatitis C disease.

5.1. LR

LR is a classification model tailored for addressing categorical classification problems, specifically those where the target variable is categorical in nature.

This method enables researchers to estimate the probability of an observation belonging to a particular class [36]. The LR classifier is based on the following logistic function:

$$p = \frac{1}{(1+e^{-x})} \tag{6}$$

During the training phase, for each instance x1, x2, x3, ..., xn, the logistic regression coefficients are denoted as b0, b1, b2, ..., bn. These coefficient values are determined and refined using the stochastic gradient descent method. Now, the coefficients values are updated by the using the following equation:

$$b = b + l * (y - p) * (1 - p) * p * x \tag{7}$$

At the outset, all coefficient values are initialized to zero, and 'y' represents the output for each training instance. 'L' denotes the learning rate, while 'x' serves as the biased input for 'b0,' consistently set to 1. The coefficients' values are iteratively updated until the model correctly predicts the output during the training stage.

5.2. RF

Random Forest is a classification algorithm based on the ensemble principle of decision trees. During the training phase, a large collection of decision trees is constructed, forming a “forest.” In the testing phase, each tree independently predicts the class label for a given data point. The final classification is determined through majority voting, where the class receiving the highest number of votes is assigned as the predicted label for that instance. This procedure is systematically applied to all data points in the dataset [37].

5.3. SVM

In the Support Vector Machine (SVM) algorithm, each data instance is represented as a point in an n-dimensional space, where n denotes the number of features. The algorithm then determines an optimal hyperplane that separates data points belonging to different classes along its boundary. Each feature in this multidimensional space corresponds to a distinct coordinate [38].

5.4. KNN

The K-Nearest Neighbors (KNN) classifier determines the class of a new data instance by identifying its k-nearest neighbors within the feature space based on the training dataset. Among various distance measures, the Euclidean distance is the most used metric for this purpose. During classification, the input sample is assigned the class label that is most frequent among its k-nearest neighbors. Before applying the KNN algorithm, it is essential to address missing values, as distance calculations cannot be accurately performed when incomplete data points are present [39].

6. Ensemble Learning (EL) Techniques

The act of enhancing predictive performance by amalgamating predictions from multiple models is termed as EL. Within the domain of EL, only three widely acknowledged approaches exist, even though various prediction modelling techniques could potentially produce numerous ensembles [40].

Types

While there are virtually countless methods to achieve this goal, it is likely that three specific types of EL approaches are the most researched and applied in practical scenarios. Their popularity is attributed to their simplicity and effectiveness in addressing a broad spectrum of predictive modelling tasks [41]. They are:

- Bagging
- Stacking
- Boosting

6.1. Boosting

In the ensemble method called "boosting," the training data is modified to emphasize instances that past models fitted to the training set might have misclassified. In the realm of boosting, dataset trained for each updated classifier is continually refined to focus on instances that important classifiers have misclassified. Rectifying prediction errors is a crucial aspect of ensemble [42].

Here are some key facets of boosting:

- Skew training data towards challenging prediction instances
- Aggregate predictions by employing a weighted average of models.

The following boosting classifiers applied for diagnosing of hepatitis C disease.

- AdaBoost
- CatBoost
- Gradient Boosting

A. AdaBoost Classifier

AdaBoost is an effective ensemble learning technique that enhances model performance by iteratively adjusting the weights of training instances to focus on difficult-to-classify examples [43]. In this study, the implementation of AdaBoost, an Adaptive Boosting approach, demonstrated the strength of boosting as a robust ensemble classification method. The boosting process resulted in improved predictive accuracy across the evaluated models [44].

B. CatBoost Classifier

CatBoost is a gradient boosting algorithm based on decision trees, specifically designed to efficiently handle categorical features. It automatically converts categorical variables into numerical representations, thereby eliminating the need for manual encoding. This inherent capability simplifies hyperparameter tuning, reduces overfitting, and improves model generalization. In this study, the CatBoost model was configured with a random strength parameter of 0.1 and a fixed number of 1000 iterations during training [45].

C. Gradient Boosting Classifier

Gradient Boosting is a robust ensemble learning technique applicable to both classification and regression problems. It constructs models sequentially, with each subsequent model aiming to correct the errors made by its predecessors. By employing gradient descent to minimize the loss function, this approach progressively enhances the performance of weak learners, culminating in a strong and highly accurate predictive model [46].

7. Class Balance Techniques

7.1. SMOTE (Synthetic Minority Over-sampling Technique)

A resampling model is introduced, in which underrepresented categories are amplified through duplication to produce artificial samples. This approach generates synthetic samples for underrepresented data points by constructing new instances between them and their k nearest neighbors within the same class [47].

The percentage of hepatitis patients with a fatal outcome is 87.80%, whereas the proportion with a survival outcome is 12.20%. To address the skewed learning behavior resulting from class imbalance, SMOTE can be used to balance the dataset1. After applying the SMOTE technique for class balancing, 384 patients died and 384 patients survived.

In the updated analysis, 79.35% of the subjects in dataset 2 were identified as *suspected hepatitis patients*, while only 20.65% were classified as *healthy*. This significant disparity between class frequencies can result in skewed model

learning, with a tendency to overrepresent the dominant class. To mitigate this problem, SMOTE was applied to achieve class balance. After applying SMOTE, the dataset included 54 suspected patients and 54 healthy individuals, thereby ensuring a more balanced representation for improved model performance.
After SMOTE class balancing technique classification techniques like LR, RF, SVM, KNN, and EL techniques were used to diagnose hepatitis C disease.

7.2. ADASYN (Adaptive Synthetic Sampling)

The ADASYN technique functions by adaptively generating artificial instances for minority-class examples based on their corresponding distributions. Specifically, it generates a larger amount of synthetic data samples for minority class instances that are more difficult to learn, compared to those minority samples that are relatively easy to learn. The ADASYN approach has the capability of reducing the learning bias that arises from the initial imbalanced distribution of data. Additionally, it can dynamically adjust the decision boundary to prioritize the challenging samples that are tough to learn [48]. After applying the ADASYN technique for class balancing, 384 patients died and 380 patients survived. After ADASYN class balancing technique classification techniques like LR, RF, SVM, KNN and EL techniques were used to diagnose hepatitis C disease. In the updated analysis, 79.35% of the subjects in dataset 2 were identified as *suspected hepatitis patients*, while only 20.65% were classified as *healthy*. This significant class imbalance can result in biased model training that favors the majority class. To overcome this issue, the ADASYN technique was applied to balance the dataset. Following the application of ADASYN, the dataset was adjusted to include 54 suspected patients and 54 healthy individuals, ensuring a more balanced class distribution and improving the model's performance on both categories.

To address data imbalance, an additional analysis was performed to assess the distribution of severity levels within the dataset. Cases were classified according to clinically relevant indicators, including elevated bilirubin levels and the presence of liver cirrhosis. The results revealed that severe cases represented a minority class, indicating a pronounced severity imbalance. To mitigate this issue, two oversampling techniques, ADASYN and SMOTE, were applied to balance the dataset and enhance the representation of the minority (severe) class during model training.

8. Performance Evaluation

The model's performance was evaluated using key metrics derived from classification outcomes. These included true positives (TP) for correctly identified positive cases, true negatives (TN) for correctly identified negative cases, false positives (FP) for incorrect positive predictions, and false negatives (FN) for missed positive instances [49].

Accuracy: n classification tasks, accuracy is defined as the ratio of correctly predicted instances to the total number of instances assessed.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{8}$$

Precision: In classification tasks, it refers to the proportion of correctly identified positive instances out of all instances that were predicted to be positive.

$$Precision = \frac{TP}{TP+FP} \tag{9}$$

Sensitivity or Recall: Recall for a given label is calculated as the proportion of true positive cases out of all actual positive instances.

$$Sensitivity = \frac{TP}{TP+FN} \tag{10}$$

Specificity: Specificity measures the system's ability to correctly identify negative cases.

$$Specificity = \frac{TN}{TN+FP} * 100 \tag{11}$$

F1- Score: The F-measure, or F1 score, is the harmonic mean of precision and sensitivity, offering a balanced evaluation of a model's ability to accurately detect relevant cases while reducing false positives.

$$F1 - SCORE = 2 * \frac{Precsion*Recall}{Precsion+Recall} \tag{12}$$

To address this issue, misclassified instances were analyzed across both ADASYN- and SMOTE-balanced datasets. The analysis revealed that most misclassifications occurred near class boundaries, particularly in logistic regression and KNN models, where decision margins tend to be less adaptable. For example, the KNN classifier demonstrated lower sensitivity (55.17% with ADASYN and 58.62% with SMOTE), highlighting challenges in correctly identifying positive

cases due to overlapping feature distributions or the sparsity of minority-class samples in the feature space. Additionally, misclassifications were more common among borderline or ambiguous cases where clinical indicators, such as ALT and AST levels, were close to diagnostic thresholds. In contrast, ensemble methods like Random Forest and Gradient Boosting achieved higher ROC-AUC and F1-scores, reflecting enhanced robustness and superior handling of borderline instances.

To promote model transparency and foster trust among healthcare professionals and patients, Explainable AI (XAI)techniques were integrated into the study. Specifically, local explanation methods such as SHAP force plots were employed to visualize the contribution of individual features to each prediction generated by the ensemble models.

9. Results and Discussion

In this study, four machine learning classifiers along with EL models were utilized for the diagnosis of hepatitis disease. Classification Models and EL techniques experiments without class balancing run on hepatitis C dataset.

Firstly, classification algorithms & EL techniques were applied with one of class balancing technique ADASYN. Next, classification algorithms & EL techniques were applied with one of class balancing technique SMOTE.

The table 4 includes the values of performance classifier after applying ADASYN class balance technique on dataset 1.LR achieved 88.11% accuracy,58.54% precision ,82.76% recall, specificity 89.10%, f1 score 68.57%, ROC-AUC Curve 0.8593. RF achieved 98.92% accuracy, 93.33% precision, 96.55% recall, specificity 98.72%, f1 score 94.92%, ROC-AUC Curve 0.9763.

SVM achieved 92.97% accuracy, 75.00% precision, 82.76% recall, specificity 94.87%, f1 score 78.69%, ROC-AUC Curve 0.888. KNN achieved 88.65% accuracy, 66.67% precision, 82.76% recall, specificity 94.87%, f1 score 60.38%, ROC-AUC Curve 0.7502. AdaBoost achieved 97.30% accuracy ,85.29% precision, recall 100.00%, specificity 96.79%, f1-score 92.06%, ROC-AUC curve 0. 9839.Catboost achieved 92.97 accuracy,71.05% precision ,93.10% recall, specificity 92.95%, f1 score 80.60%, ROC-AUC curve 0.9302. Gradient Boost achieved 96.22% accuracy, precision 80.56%, sensitivity 100.00%, specificity 95.51% f1score 89.23%, ROC-AUC 0.9775.

Table 4. Classification algorithms and E techniques performance after applying ADASYN class balance technique

Classifier	*Accuracy*	*Precision*	*Sensitivity or Recall*	*Specificity*	*F1- Score*	*ROC-AUC Curve*
LR	88.11%	58.54%	82.76%	89.10%	68.57%	0.8593
RF	98.92%	93.33%	96.55%	98.72%	94.92%	0.9763
SVM	92.97%	75.00%	82.76%	94.87%	78.69%	0.888
KNN	88.65%	66.67%	55.17%	94.87%	60.38%	0.7502
AdaBoost	97.30%	85.29%	100.00%	96.79%	92.06%	0.9839
CatBoost	92.97	71.05%	93.10%	92.95%	80.60%	0.9302
Gradient Boost	96.22%	80.56%	100.00%	95.51%	89.23%	0.9775

The performance of machine learning classifiers and ensemble models was evaluated both with and without the application of class balancing [48]. Table 5 presents the performance metrics obtained after applying the SMOTE balancing technique.LR achieved 90.27% accuracy,65.71% precision ,79.31% recall, specificity 92.31%, f1 score 71.88%, ROC-AUC Curve 0.8580. RF achieved 98.38 % accuracy, 93.33% precision, 96.55% recall, specificity 98.72%, f1 score 94.92%, ROC-AUC Curve 0.9763.

SVM achieved 94.59% accuracy, 82.76% precision, 82.76% recall, specificity 96.79%, f1 score 82.76%, ROC-AUC Curve 0.8977. KNN achieved 89.73% accuracy,70.83% precision, 58.62% recall, specificity 95.51%, f1 score 64.13%, ROC-AUC Curve 0.7706. AdaBoost achieved 96.76% accuracy ,87.10% precision, recall 93.10%, specificity 97.44%, f1-score 90.00%, ROC-AUC curve 0. 9526.Catboost achieved 95.14 accuracy,79.41% precision ,93.10% recall, specificity 95.51%, f1 score 85.71%, ROC-AUC curve 0.9430. Gradient Boost achieved 96.76% accuracy, precision 84.85%, sensitivity 96.55%, specificity 96.79% f1score 90.32%, ROC-AUC 0.9667.

Table 5. Classification algorithms and EL techniques performance after applying SMOTE class balance technique.

Classifier	*Accuracy*	*Precision*	*Sensitivity or Recall*	*Specificity*	*F1- Score*	*ROC-AUC Curve*
LR	90.27%	65.71%	79.31%	92.31%	71.88%	0.8580
RF	**98.38%**	93.33%	96.55%	98.72%	94.92%	0.9763
SVM	94.59%	82.76%	82.76%	96.79%	82.76%	0.8977
KNN	89.73%	70.83%	58.62%	95.51%	64.15%	0.7706
AdaBoost	96.76%	87.10%	93.10%	97.44%	90.00%	0.9526
CatBoost	95.14%	79.41%	93.10%	95.51%	85.71%	0.9430
Gradient Boost	96.76%	84.85%	96.55%	96.79%	90.32%	0.9667

The proposed model achieved a training accuracy of 100% and a testing accuracy of 97.56%, reflecting a minimal generalization gap of only 2.44%. This small gap suggests that the model generalizes well to unseen data and does not exhibit significant overfitting. The table 6 summarizes the performance of various classification and EL algorithms after applying the SMOTE technique to balance the dataset 2. Among all models, Random Forest achieved the highest accuracy (93.75%) and perfect recall (100%), while AdaBoost showed excellent performance with perfect precision (1.0) and a high F1-score (91.66%). KNN also performed well, especially in identifying healthy patients with perfect specificity. In contrast, SVM and CatBoost showed relatively lower recall, indicating limitations in detecting suspected cases. Overall, ensemble models like Random Forest and AdaBoost outperformed standalone classifiers in handling the balanced dataset effectively.

Table 6. Classification algorithms and EL techniques performance after applying SMOTE class balance technique

Classifier	Accuracy	Precision	Sensitivity or Recall	Specificity	F1- Score
LR	81.25	91.66	84.61	66.66	88.00
RF	93.75	92.85	100.00	66.66	96.29
SVM	75.00	90.90	76.92	66.66	83.33
KNN	87.25	91.66	84.60	1.0	91.66
AdaBoost	87.50	1.0	84.61	1.0	91.66
CatBoost	75.50	1.0	69.23	1.0	81.81
Gradient Boost	81.25	91.66	84.61	66.66	88.00

Table 7 summarizes the performance of different classifiers after applying the SMOTE technique to balance Dataset 2. Among all models, AdaBoost delivered the best results, achieving the highest accuracy (93.75%), perfect precision (100%), and the highest ROC-AUC score (0.96). Random Forest also demonstrated strong performance, maintaining balanced precision and recall (92.31%) with an overall accuracy of 87.50%. Although the SVM classifier achieved perfect recall, its 0% specificity and low ROC-AUC (0.50) reflect poor differentiation of healthy cases. KNN, CatBoost, and Gradient Boosting exhibited moderate performance levels, while Logistic Regression showed commendable precision and specificity. Overall, AdaBoost and Random Forest emerged as the most effective models on the SMOTE-balanced dataset.

Table 7. Classification algorithms and EL techniques performance after applying ADASYN class balance technique

Classifier	**Accuracy**	**Precision**	**Sensitivity or Recall**	**Specificity**	**F1- Score**	**ROC-AUC Curve**
LR	81.25%	100.00%	76.92%	100.00%	86.96%	0.88
RF	87.50%	92.31%	92.31%	66.67%	92.31%	0.79
SVM	68.75%	81.25%	100.00%	0.00%	89.66%	0.50
KNN	81.25%	85.71%	92.31%	33.33%	88.89%	0.62
AdaBoost	**93.75%**	100.00%	92.31%	100.00%	96.00%	0.96
CatBoost	81.25%	85.71%	92.31%	33.33%	88.89%	0.62
Gradient Boost	81.25%	91.67%	84.62%	66.67%	88.00%	0.75

Figure 13 illustrates the AUC curve without class imbalance for all ML Models and EL techniques. Here CatBoost classifier has highest AUC value 1.00.

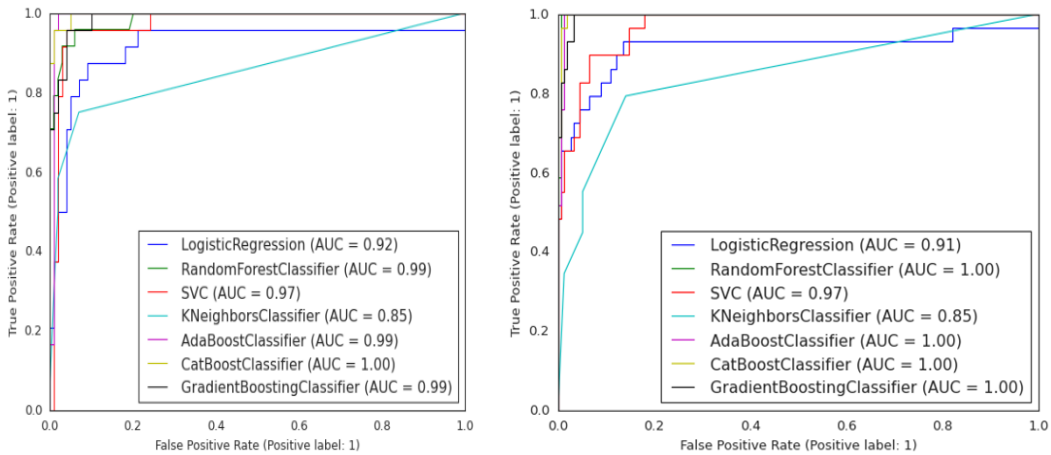


Fig.13. ROC curve for Models without class imbalance and with ADASYN class balance Technique

The following Figure 14 illustrates the AUC curve with SMOTE class balance technique. Here RF, CatBoost, Gradient Boost has the highest AUC score 1.00.

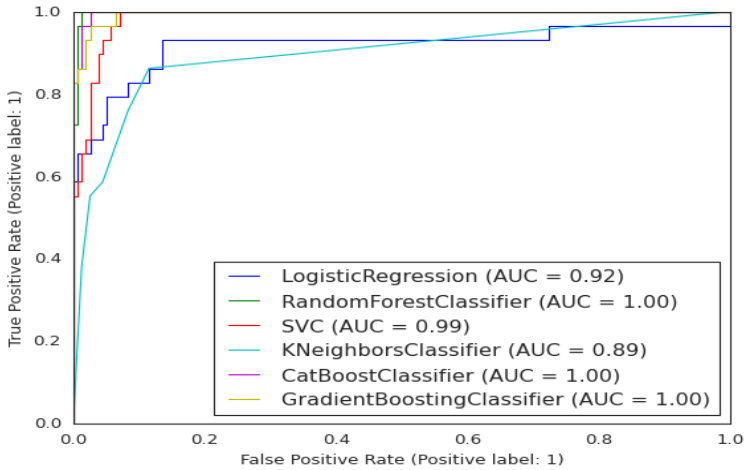


Fig.14. ROC curve for the models with SMOTE class balance technique

Figure 15 presents the ROC curves showing the performance of various classification models after applying SMOTE and ADASYN techniques to address class imbalance in Dataset 1. With SMOTE, AdaBoost and CatBoost achieved perfect classification results (AUC = 1.0), while Random Forest, KNN, and Gradient Boosting also performed strongly with AUC scores of 0.97. Logistic Regression achieved a good score of 0.92, and Support Vector Machine showed the lowest AUC of 0.87. Under ADASYN, Naive Bayes and AdaBoost yielded the highest AUC values (0.97), followed by Logistic Regression and Random Forest (0.92), CatBoost (0.95), and Gradient Boosting (0.90). SVM and KNN had lower AUC scores of 0.87 and 0.81, respectively. Overall, ensemble models consistently outperformed others across both sampling techniques, and the use of SMOTE and ADASYN effectively improved classification performance by addressing the issue of class imbalance. The Figure 16 illustrates the accuracy comparison of ML Classifiers and EL Classifiers on unbalanced, ADASYN, SMOTE Balancing Techniques. AdaBoost without class balancing technique was 97.56%.

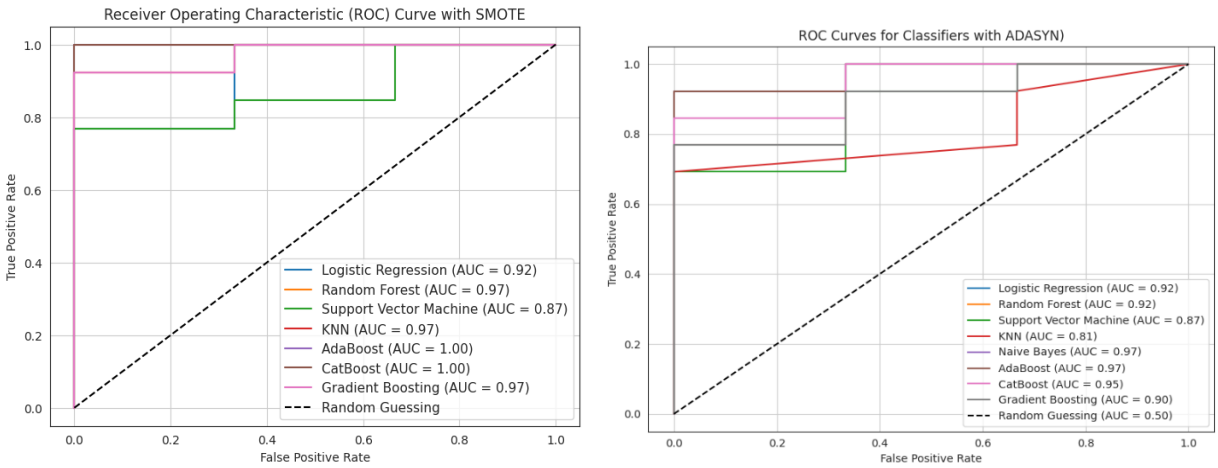


Fig.15. ROC curve for the models with SMOTE & ADASYN class balance technique

Figure 16 displays the evaluation metrics, precision, recall, F1 score, and specificity, used to evaluate and contrast the effectiveness of machine learning and ensemble-based models. In terms of F1 score, Random Forest (RF) achieved the highest score. For precision, both CatBoost and Gradient Boost exhibited the highest scores. In the case of Recall, both CatBoost and Gradient Boost obtained the highest scores. Specificity, on the other hand, showed the same highest score for RF, CatBoost, and Gradient Boost.

Figure 17 presents the evaluation metrics used to analyze and differentiate the effectiveness of ML models and EL techniques applied using the ADASYN technique. In terms of F1 score, Adaboost achieved the highest score. For precision, both AdaBoost exhibited the highest scores. In the case of Recall, RF, AdaBoost, CatBoost and Gradient Boost obtained the highest scores. Specificity, on the other hand, showed the RF obtained the highest score.

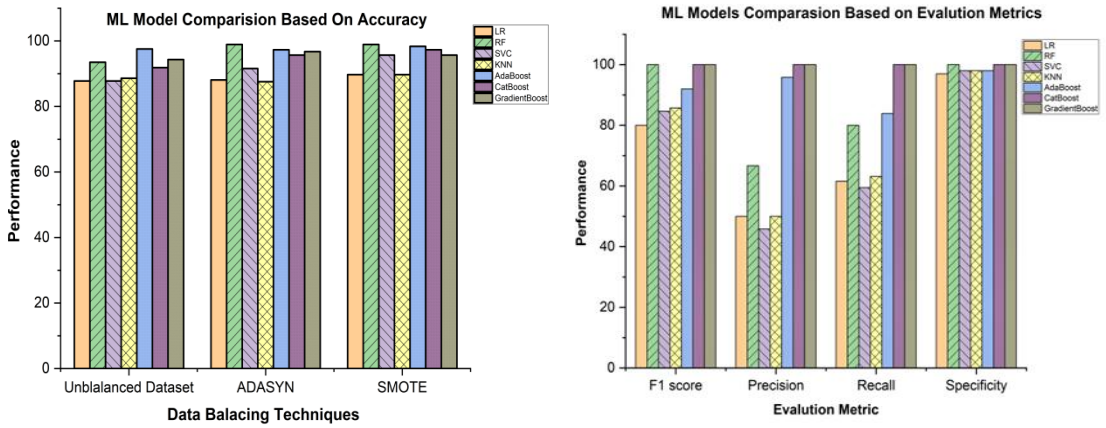


Fig.16. The Accuracy & Evaluation metrics comparison of the ML Models on unbalanced Dataset, ADASYN, SMOTE

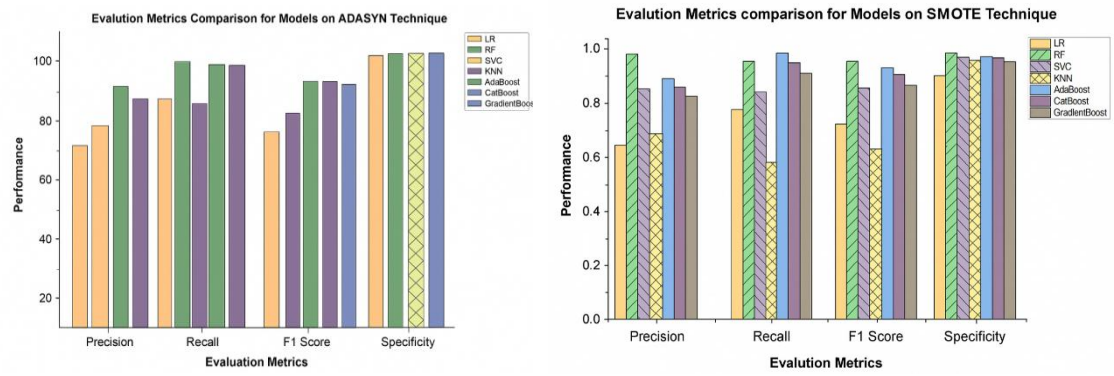


Fig.17. Evaluation metrics after applying ADASYN & SMOTE balancing technique

Figure 17 illustrates the evaluation metrics used to assess and contrast the effectiveness of various ML models and EL methods applied with the SMOTE technique. In terms of F1 score, Adaboost achieved the highest score. For precision, RF exhibited the highest scores. In the case of Recall AdaBoost obtained the highest score. Specificity, on the other hand, showed the RF obtained the highest score.

The findings of this study contribute to the theoretical understanding of how oversampling techniques such as ADASYN and SMOTE affect classifier performance, particularly regarding F1-score and specificity. These results extend existing frameworks for managing imbalanced datasets. From a practical standpoint, the insights gained facilitate informed decision-making in critical areas like medical diagnostics, where choosing an appropriate oversampling strategy and classifier is essential to minimizing the impact of misclassification.

Computational Efficiency

To assess the real-world, deploy ability of the models, computational metrics such as training time, prediction latency, and model size were analyzed are specified in below table 8. The following table 9 summarize these metrics, highlighting each model's suitability for deployment in resource-constrained environments.

Table 8. Comparison of machine learning models based on computational efficiency with ADASYN balancing technique

Sno	Model	Training Time (s)	Prediction Latency (s/sample)	Model Size (KB)
1	Logistic Regression	0.0330	0.000004	1.09
2	Random Forest Classifier	0.5670	0.000090	431.27
3	SVM	0.1315	0.000139	67.12
4	KNN Classifier	0.0011	0.000015	139.80
5	Ada Boost Classifier	1.1297	0.000216	63.91
6	CatBoost Classifier	17.4435	0.000093	1103.11
7	Gradien Boosting Classifier	1.6903	0.000009	137.77

Table 9. Comparison of machine learning models based on computational efficiency with SMOTE balancing technique

Sno	Model	Training Time (s)	Prediction Latency (s/sample)	Model Size (KB)
1	Logistic Regression	0.0096	0.000006	1.09
2	Random Forest Classifier	0.2894	0.000095	360.96
3	SVM	0.0390	0.000019	46.04
4	KNN Classifier	0.0010	0.000014	139.32
5	Ada Boost Classifier	0.5212	0.000294	63.91
6	CatBoost Classifier	11.1582	0.000020	1100.29
7	Gradient Boosting Classifier	0.5382	0.000007	136.65

10. Conclusion & Future Scope

In this work, a ML based approach was developed for the diagnosis of hepatitis using both classification algorithms and EL techniques. These models were evaluated on two Hepatitis C datasets, both before and after applying class balancing techniques. The implemented classifiers included Logistic Regression, Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN), along with ensemble methods such as AdaBoost, CatBoost, and Gradient Boosting. The application of the ADASYN technique significantly improved classification performance, with Random Forest achieving an accuracy of 98.38%, followed by AdaBoost and Gradient Boosting at 96.76%, and SVM at 92.43%. Similarly, the use of SMOTE further enhanced model performance, with AdaBoost and Random Forest again reaching the highest accuracy of 98.38%, and SVM improving to 94.05%.

For Dataset 2, SMOTE enabled Random Forest to achieve the highest accuracy (93.75%) and perfect recall (100%), while ADASYN boosted AdaBoost's performance, resulting in 93.75% accuracy, 100% precision, and the highest ROC-AUC score (0.96). KNN and Gradient Boosting also demonstrated strong results across both balancing methods. Although SVM achieved high recall with ADASYN, its specificity and ROC-AUC remained relatively low. Overall, the findings highlight that EL techniques, particularly AdaBoost and Random Forest, consistently outperformed individual classifiers in the presence of class imbalance, making them highly effective for hepatitis classification tasks.

Future research directions include the integration of hybrid oversampling techniques and the exploration of cost-sensitive learning frameworks to enhance classification performance, particularly on imbalanced clinical datasets. To further ensure the model's generalizability and robustness, future research will focus on validating the proposed models using independent datasets collected from diverse geographic regions and clinical settings. In response to current challenges, the study has also been expanded to incorporate a stacking ensemble framework that integrates the predictive capabilities of multiple base classifiers, aiming to enhance overall model accuracy and robustness. Deployment strategies will involve exposing the models as RESTful APIs or embedding them into systems compatible with existing Electronic Health Record (EHR) platforms and point-of-care devices.

Furthermore, external validation using datasets from diverse regions and healthcare systems is planned to evaluate the model's robustness, adaptability, and generalizability. This step is essential for ensuring clinical applicability across varied populations and for mitigating potential biases. Interoperability with clinical systems will be further supported through adherence to standards such as HL7 FHIR, enabling seamless data exchange and alignment with routine clinical practices Robustness of the predictive models will also be confirmed through extensive cross-validation and evaluation using metrics such as AUC and F1-score.

Regarding ethical considerations, this study utilized publicly available, anonymized datasets in full compliance with relevant data-sharing regulations, ensuring that no personally identifiable information (PII) was included. The original dataset providers obtained the necessary informed consent and ethical approvals prior to data collection. For future clinical deployment, securing patient consent and maintaining transparent communication will be crucial. Furthermore, the study emphasizes the importance of equitable access, particularly in underdeveloped or resource-limited regions. The lightweight and scalable design of the proposed model supports its deployment in such settings, thereby fostering inclusive and accessible AI-driven healthcare solutions.

Author Contributions Statement

Karthika Natarjan– Conceptualization, methodology, supervision, review, and editing: Proposed the research ideas, constructed the overall framework, supervised project execution, reviewed and edited the manuscript, ensured clarity and coherence, and helped coordinate project milestones and deadlines.

Koteswara Rao Makkena – Data curation, software implementation, formal analysis, visualization, statistical analysis, and writing: Handled data acquisition and preprocessing, implemented the research model, led model training, validation, and performance benchmarking against existing methods, performed in-depth analysis of experimental results, prepared performance charts and visualizations, ensured the statistical robustness of the evaluation, and drafted the initial manuscript, including the literature survey and technical background of the study.

Conflict of Interest Statement
The authors declare no conflicts of interest.

Funding Declaration
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors

Data Availability Statement
This study analyzed publicly available Hepatitis C datasets obtained from Kaggle and the UCI Machine Learning Repository.

Ethical Declarations
This study did not involve the prospective recruitment of human participants or experiments on animals. It was conducted exclusively on publicly available, fully anonymized secondary datasets obtained from Kaggle, for which the original data providers had already secured the necessary informed consent and ethical approvals. Therefore, according to the institutional and national guidelines applicable to this work, no additional approval from a human or animal ethics committee was required.

Acknowledgments
We sincerely thank the experts for their professional evaluation and valuable recommendations, which have contributed to improving the quality of the experiment and the reliability of its results.

Declaration of Generative AI in Scholarly Writing
Generative AI and AI-assisted tools were used only to improve the language and readability of this manuscript, such as for grammar checking, wording suggestions, and formatting of sentences. All scientific content, including the conception of the study, data analysis, interpretation of results, and conclusions, was developed and verified by the authors, who carefully reviewed and edited all AI-assisted text. No generative AI system is listed as an author or co-author, and the authors accept full responsibility for the accuracy and integrity of the content presented in this article.

Abbreviations
The following abbreviations are used in this manuscript:

ADASYN – Adaptive Synthetic Sampling
AIC – Akaike Information Criterion (if present in your figures/tables; not in text snippet)
ALB – Albumin
ALP – Alkaline Phosphatase
ALT – Alanine Aminotransferase
ANN – Artificial Neural Network
AST – Aspartate Aminotransferase
AUC – Area Under the ROC Curve
C4.5 – C4.5 Decision Tree Algorithm
CatBoost – Categorical Boosting (gradient boosting algorithm)
CHE – Cholinesterase
CHOL – Cholesterol
DT – Decision Tree
EHR – Electronic Health Record
EL – Ensemble Learning
FN – False Negative
FP – False Positive
GGT – Gamma-Glutamyl Transferase
HL7 FHIR – Health Level Seven Fast Healthcare Interoperability Resources
ID – Identifier (patient ID in dataset)
KNN – K-Nearest Neighbors
LIME – Local Interpretable Model-Agnostic Explanations
LightGBM – Light Gradient Boosting Machine
LR – Logistic Regression
MCC – Matthews Correlation Coefficient (in cited work)
MeitY – Ministry of Electronics and Information Technology
ML – Machine Learning
MLP – Multilayer Perceptron

NB – Naive Bayes
PIDD – Pima Indians Diabetes Dataset
PROT – Total Protein
RF – Random Forest
ROC – Receiver Operating Characteristic
SFS – Sequential Forward Selection
SGOT / AST – Serum Glutamic Oxaloacetic Transaminase / Aspartate Aminotransferase
SHAP – SHapley Additive exPlanations
SMOTE – Synthetic Minority Over-sampling Technique
SVM – Support Vector Machine
TP – True Positive
TN – True Negative
TPR – True Positive Rate
UCI – University of California Irvine (Machine Learning Repository)
UL – Units per Liter
WHO – World Health Organization
XAI – Explainable Artificial Intelligence
XGBoost – Extreme Gradient Boosting

References

[1] K. Polat and S. Güneş, "Prediction of hepatitis disease based on principal component analysis and artificial immune recognition system," *Appl. Math. Comput.*, vol. 189, no. 2, pp. 1282–1291, 2007, doi:10.1016/j.amc.2006.12.010.

[2] E. Dogantekin, A. Dogantekin, and D. Avci, "Automatic hepatitis diagnosis system based on linear discriminant analysis and adaptive network based fuzzy inference system," *Expert Syst. Appl.*, vol. 36, no. 8, pp. 11282–11286, 2009, doi:10.1016/j.eswa.2008.02.033.

[3] World Health Organization, "Hepatitis C," Fact sheet, 2024. Available: https://www.who.int/news-room/fact-sheets/detail/hepatitis-c

[4] B. Rehermann and M. Nascimbeni, "Immunology of hepatitis B virus and hepatitis C virus infection," *Nat. Rev. Immunol.*, vol. 5, no. 3, pp. 215–229, 2005, doi:10.1038/nri1573.

[5] K. Polat and S. Güneş, "Hepatitis disease diagnosis using a new hybrid system based on feature selection and artificial immune recognition system with fuzzy resource allocation," *Digit. Signal Process.*, vol. 16, no. 6, pp. 889–901, 2006, doi:10.1016/j.dsp.2006.06.003.

[6] E. K. AbuSharekh and S. S. Abu-Naser, "Diagnosis of hepatitis virus using artificial neural network," *Int. J. Acad. Eng. Res.*, vol. 2, no. 11, pp. 1–7, 2018.

[7] E. A. Bayrak, P. Kirci, and T. Ensari, "Performance analysis of machine learning algorithms and feature selection methods on hepatitis disease," *Int. J. Multidiscip. Stud. Innov. Technol.*, vol. 3, no. 2, pp. 135–138, 2019.

[8] M. Nilashi, H. Ahmadi, L. Shahmoradi, O. Ibrahim, and E. Akbari, "A predictive method for hepatitis disease diagnosis using ensembles of neuro-fuzzy technique," *J. Infect. Public Health*, vol. 12, no. 1, pp. 13–20, 2019, doi:10.1016/j.jiph.2018.09.008.

[9] World Health Organization Regional Office for Africa, "Prevention, care and treatment of viral hepatitis in the African region: framework for action, 2016–2020," 2017. Available: https://www.afro.who.int/publications/prevention-care-and-treatment-viral-hepatitis-african-region-framework-action-2016

[10] C. A. U. Hassan, M. S. Khan, and M. A. Shah, "Comparison of machine learning algorithms in data classification," in *Proc. 24th Int. Conf. Autom. Comput. (ICAC)*, 2018, pp. 1–6, doi:10.23919/IConAC.2018.8748993.

[11] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi:10.1023/A:1010933404324.

[12] V. Vapnik and A. Chervonenkis, "The necessary and sufficient conditions for consistency in the empirical risk minimization method," *Pattern Recognit. Image Anal.*, vol. 1, no. 3, pp. 283–305, 1991, doi:10.1007/BF00941892.

[13] C. A. U. Hassan, M. S. Khan, and M. A. Shah, "Comparison of machine learning algorithms in data classification," in *Proc. 24th Int. Conf. Autom. Comput. (ICAC)*, 2018, pp. 1–6.

[14] W. Akbar, W. P. Wu, S. Saleem, M. Farhan, M. A. Saleem, A. Javeed, and L. Ali, "Development of hepatitis disease detection system by exploiting sparsity in linear support vector machine to improve strength of AdaBoost ensemble model," *Mobile Inf. Syst.*, vol. 2020, Art. ID 8870240, pp. 1–9, 2020, doi:10.1155/2020/8870240.

[15] K. Kanagarathinam, D. Sankaran, and R. Manikandan, "Machine learning-based risk prediction model for cardiovascular disease using a hybrid dataset," *Data Knowl. Eng.*, vol. 140, Art. ID 102042, 2022, doi:10.1016/j.datak.2022.102042.

[16] H. Park *et al.*, "Machine learning algorithms for predicting direct-acting antiviral treatment failure in chronic hepatitis C: An HCV-TARGET analysis," *Hepatology*, vol. 76, no. 2, pp. 483–491, 2022, doi:10.1002/hep.32316.

[17] Kaggle, "Hepatitis C predictions – HepatitisCdata.csv," 2019. Available: https://www.kaggle.com/code/krishnabhatt4/hepatitis-c-predictions/input/HepatitisCdata.csv

[18] D. Dua and C. Graff, "Hepatitis Data Set," UCI Machine Learning Repository, Univ. California, Irvine, CA, USA, 2019. Available: https://archive.ics.uci.edu/ml/datasets/hepatitis, doi:10.24432/C5JP5K.

[19] X. Tian *et al.*, "Using machine learning algorithms to predict hepatitis B surface antigen seroclearance," *Comput. Math. Methods Med.*, vol. 2019, Art. ID 8950193, pp. 1–9, 2019, doi:10.1155/2019/8950193.

[20] K. K. Napa and V. Dhamodaran, "Hepatitis-infectious disease prediction using classification algorithms," *Res. J. Pharm. Technol.*, vol. 12, no. 8, pp. 3720–3725, 2019, doi:10.5958/0974-360X.2019.00639.4.

[21] V. K. Yarasuri, G. K. Indukuri, and A. K. Nair, "Prediction of hepatitis disease using machine learning technique," in *Proc. 3rd Int. Conf. I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, 2019, pp. 265–269, doi:10.1109/I-SMAC47947.2019.9032542.
[22] A. Orooji and F. Kermani, "Machine learning based methods for handling imbalanced data in hepatitis diagnosis," *Front. Health Inform.*, vol. 10, no. 1, p. 57, 2021, doi:10.30699/fhi.v10i1.57.
[23] M. Glučina, A. Lorencin, N. Anđelić, and I. Lorencin, "Cervical cancer diagnostics using machine learning algorithms and class balancing techniques," *Appl. Sci.*, vol. 13, no. 2, p. 1061, 2023, doi:10.3390/app13021061.
[24] K. Sachdeva, P. Bathla, P. Rani, V. Solanki, and R. Ahuja, "A systematic method for diagnosis of hepatitis disease using machine learning," *Innov. Syst. Softw. Eng.*, vol. 19, no. 1, pp. 71–80, 2023, doi:10.1007/s11334-022-00463-2.
[25] A. Doğru, S. Buyrukoğlu, and M. Arı, "A hybrid super EL model for the early-stage prediction of diabetes risk," *Med. Biol. Eng. Comput.*, vol. 61, no. 3, pp. 785–797, 2023, doi:10.1007/s11517-022-02719-7.
[26] N. H. Taz, A. Islam, and I. Mahmud, "A comparative analysis of ensemble-based machine learning techniques for diabetes identification," in *Proc. 2nd Int. Conf. Robot., Electr. Signal Process. Techn. (ICREST)*, 2021, pp. 1–6, doi:10.1109/ICREST51555.2021.9331154.
[27] Z. Sadeghi *et al.*, "A brief review of explainable artificial intelligence in healthcare," *arXiv* preprint arXiv:2304.01543, 2023.
[28] A. M. Ali *et al.*, "Explainable machine learning approach for hepatitis C diagnosis using SFS feature selection," *Machines*, vol. 11, no. 3, p. 391, 2023, doi:10.3390/machines11030391.
[29] K. D. Mandl *et al.*, "SMART on FHIR: A standards-based, interoperable apps platform for electronic health records," *J. Am. Med. Inform. Assoc.*, vol. 23, no. 5, pp. 899–908, 2016, doi:10.1093/jamia/ocv189.
[30] H. Lin, R. Zhang, and T. Tong, "When Tukey meets Chauvenet: A new boxplot criterion for outlier detection," *arXiv* preprint arXiv:2506.XXXXX, 2025. (Update arXiv ID when available.)
[31] S. Prasad *et al.*, "Feature selection using ANOVA F-test with SelectKBest in load recognition for home energy management systems," *Sensors*, vol. 24, no. 15, p. 4965, 2024, doi:10.3390/s24154965.
[32] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 4, pp. 491–502, 2005, doi:10.1109/TKDE.2005.66.
[33] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 785–794, doi:10.1145/2939672.2939785.
[34] D. Sinha, "Using and scaling SHAP on Databricks for turbocharging your model explainability: Part 1," *Medium*, 2023. Available: https://medium.com/@debusinha2009/using-and-scaling-shap-on-databricks-for-turbocharging-your-model-explainability-part-1-29099006f6b4
[35] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD '16)*, 2016, pp. 1135–1144, doi:10.1145/2939672.2939778.
[36] W. Książek, M. Gandor, and P. Pławiak, "Comparison of various approaches to combine logistic regression with genetic algorithms in survival prediction of hepatocellular carcinoma," *Comput. Biol. Med.*, vol. 134, p. 104431, 2021, doi:10.1016/j.compbiomed.2021.104431.
[37] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi:10.1023/A:1010933404324.
[38] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995, doi:10.1007/BF00994018.
[39] R. K. Halder, M. N. Uddin, M. A. Uddin, S. Aryal, and A. Khraisat, "Enhancing K-nearest neighbor algorithm: A comprehensive review and performance analysis of modifications," *J. Big Data*, vol. 11, p. 113, 2024, doi:10.1186/s40537-024-00863-4.
[40] I. D. Mienye and Y. Sun, "A survey of EL: Concepts, algorithms, applications, and prospects," *IEEE Access*, vol. 10, pp. 99129–99149, 2022, doi:10.1109/ACCESS.2022.3203283.
[41] M. A. Mohammed, B. Al-Khateeb, M. K. Abd Ghani, S. A. Mostafa, and M. S. Maashi, "A comprehensive review on ensemble deep learning: Opportunities and challenges," *J. King Saud Univ. – Comput. Inf. Sci.*, vol. 35, no. 2, pp. 757–774, 2023, doi:10.1016/j.jksuci.2021.05.012.
[42] I. D. Mienye and Y. Sun, "A survey of EL: Concepts, algorithms, applications, and prospects," *IEEE Access*, vol. 10, pp. 99129–99149, 2022, doi: 10.1109/ACCESS.2022.3207287. (Duplicate of, keep only one in final list.)
[43] W. Chen, K. Yang, Z. Yu *et al.*, "A survey on imbalanced learning: Latest research, applications and future directions," *Artif. Intell. Rev.*, vol. 57, p. 137, 2024, doi:10.1007/s10462-024-10706-y.
[44] DataCamp, "AdaBoost classifier in Python," 2023. Available: https://www.datacamp.com/tutorial/adaboost-classifier-python
[45] G. Huang *et al.*, "Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions," *J. Hydrol.*, vol. 574, pp. 1029–1041, 2019, doi:10.1016/j.jhydrol.2019.04.059.
[46] A. Doğru, S. Buyrukoğlu, and M. Arı, "A hybrid super EL model for the early-stage prediction of diabetes risk," *Med. Biol. Eng. Comput.*, vol. 61, no. 3, pp. 785–797, 2023.
[47] N. V. Chawla *et al.*, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002, doi:10.1613/jair.953.
[48] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IJCNN)*, 2008, pp. 1322–1328, doi:10.1109/IJCNN.2008.4633969.
[49] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009, doi:10.1016/j.ipm.2009.03.002.

Authors' Profiles



Koteswara Rao Makkena earned a Bachelor of Science in Computer Science from Andhra Christian College, ANU, Guntur, Andhra Pradesh, India, in 2005. He completed a Master of Computer Applications from RVR \& JC College of Engineering, ANU, Guntur, Andhra Pradesh, in 2008, and later obtained a Master of Technology in Computer Science and Engineering from NIET, JNTUK University, Kakinada, Andhra Pradesh, in 2012. Currently, he is a Research Scholar at the School of Computer Engineering, VIT-AP University, Amaravati. With four years of experience in education and research, his primary research interests include leveraging advanced AI techniques in machine learning and deep learning to predict liver diseases.



Dr. Karthika Natarajan earned her Bachelor of Engineering (BE) and Master of Engineering (ME) degrees from Anna University, followed by advanced research at the National Institute of Technology, Trichy. She has served as a Research Fellow on a project sponsored by the Ministry of Electronics and Information Technology (MeitY). Her primary research interests include information retrieval, machine learning, and deep learning. Dr. Natarajan has published her work in renowned journals and conferences and is an active member of the Institute of Electrical and Electronics Engineers (IEEE).

How to cite this paper: Karthika Natarajan, Koteswara Rao Makkena, "Hepatitis C Diagnosis using Supervised Machine Learning Algorithms and Ensemble Learning Techniques", International Journal of Information Technology and Computer Science(IJITCS), Vol.18, No.2, pp.161-182, 2026. DOI:10.5815/ijitcs.2026.02.10