EXPLAINABLE ARTIFICIAL INTELLIGENCE-BASED DIAGNOSIS ASSISTANT OF HEPATITIS C VIRUS

A. Oyedeji [1]*
M. Osifeko [1]
S. Iyiade [1]
T. Oke [1]
A. Olayiwola [1]

Received: 23/02/2025
Revised: 08/04/2025
Accepted: 09/04/2025

[1] Department of Computer Engineering, Olabisi Onabanjo University, Ago-Iwoye, Nigeria:
Corresponding Author's Email: oyedeji.ajibola@oouagoiwoye.edu.ng

Explainable Artificial Intelligence-Based Diagnosis Assistant of Hepatitis C Virus

A. Oyedeji
Department of Computer Engineering,
Olabisi Onabanjo University,
Ago-Iwoye, Nigeria
oyedeji.ajibola@oouagoiwoye.edu.ng

M. Osifeko
Department of Computer Engineering,
Olabisi Onabanjo University,
Ago-Iwoye, Nigeria
osifeko.martins@oouagoiwoye.edu.ng

S. Iyiade
Department of Computer Engineering,
Olabisi Onabanjo University,
Ago-Iwoye, Nigeria
omosubomiiyiade@gmail.com

T. Oke
Department of Computer Engineering,
Olabisi Onabanjo University,
Ago-Iwoye, Nigeria
okeracheal911@gmail.com

A. Olayiwola
Department of Computer Engineering,
Olabisi Onabanjo University,
Ago-Iwoye, Nigeria
olayiwola.abisola@oouagoiwoye.edu.ng

Abstract— Hepatitis C is a liver infection prevalent in developing countries, and early detection of this disease would significantly reduce the mortality rate. Advances in artificial intelligence have led to the development of medical diagnostics systems. However, the decisions gotten from these systems are not easily explainable to the end users. Data preprocessing, including feature scaling and oversampling using Synthetic Minority Oversampling Technique, was carried out on HCV data. Seven classifiers—logistic regression, decision tree, random forest, support vector machine, gradient boosting, K-nearest neighbor, and multilayer perceptron (MLP)—were implemented. The models were evaluated, and Shapley Additive Explanations (SHAP) values were employed for model interpretability. MLP with standard scaling has the best performance with an accuracy of 0.97 and a sensitivity and specificity of 1.00. The features with the most influence on the outcome are the albumin test, alkaline phosphatase, alanine transaminase, and aspartate aminotransferase, while sex and cholesterol had the least influence. A web-based diagnosis assistant was deployed for early diagnosis.

Keywords— Hepatitis C, Artificial Intelligence, Explainable AI, SHAP, Diagnosis Assistant.

I. INTRODUCTION

Hepatitis C is a prevalent liver infection caused by the hepatitis C virus (HCV) and is likewise a major cause of liver cancer and liver transplant. The virus can result in both acute and chronic hepatitis, with symptoms ranging from a mild infection to severe, lifelong contamination that includes malignancy, cirrhosis, and damage to the liver. Most hepatitis C infections result from exposure to infected blood through unsafe practices [1]. Approximately 1.5 million cases of the hepatitis C virus are diagnosed yearly, with about 58 million people across the globe having a persistent infection. This phenomenon is more prevalent in poorer developing nations, especially in Africa and Asia [2]. Approximately 3.2 million children and adolescents worldwide have chronic hepatitis C infection, of which the mortality is estimated to be about 290,000 individuals for the year 2019 because of cirrhosis and hepatocellular carcinoma (primary liver cancer) [1, 3].

Early discovery and diagnosis of the disease are essential for effective and timely treatment, as this can result in a reduction of morbidity and mortality rates worldwide. Delayed treatment of the disease can result in other life-threatening diseases such as cirrhosis, fibrosis, and many other complications [4]. The medical diagnosis of liver diseases most often involves a laboratory blood test to determine the levels of liver-related metabolites, more commonly referred to as the liver function tests, and hepatitis C antibody tests, which explain whether the patient is infected with HCV [4, 5, 6]. Most of these tests are not as easy to perform and thus negatively impact the early detection of the disease. The use of available, reliable, and verifiable HCV data for the diagnosis of the disease has therefore become a viable option using artificial intelligence [6].

The healthcare industry has historically been a pioneer in implementing new technologies such as the invention of new medical procedures and the management of chronic diseases [7]. Machine learning is now playing a significant part in disease prognosis and treatment [8], medical imaging and diagnostic services [9], the development of novel pharmaceuticals [10], and the management of medical records [11]. Recent research and advancements have led to the development of medical diagnostic systems utilizing the potential and decision-making ability of artificial intelligence (AI) and machine learning (ML). The remarkable progress experienced in diagnostics has been made possible by the ability of AI-based systems to analyze, break down, and comprehend information from complex data available [6, 12, 13].

As much as the performance of the model's predictions is important, the explainability and interpretability of the models are much more paramount to understanding how the conclusions have been reached. This is even more essential and expedient in data-driven healthcare AI assistants to give the end-users, particularly health workers and physicians, the necessary support to make informed decisions [14]. Understanding the basis for the diagnosis of a particular health condition based on the various inputs or features provided to the models is of great significance to reinforcing and validating the decision of the physicians. SHAP (Shapley Additive Explanations) is an efficient tool for explainable AI both for global and local model interpretability which is based on the game's theoretical Shapley values [15, 16, 17].

An analysis of the accuracy of existing ML methods and a novel AI-based ensemble model on the prediction of hepatitis C disease was the focus of the research [13]. It was observed that the proposed ensemble method had the best accuracy of 95.6%, followed by the Quick, Unbiased, Efficient, Statistical Tree and Bayesian network with accuracies of 94.6% and 94.5%, respectively. However, the study failed to consider the

sensitivity and specificity of the models to understand the kind of errors the models were making. In [18], Cascade RF-LR implemented with oversampling utilizing the artificial bee colony algorithm proposed to automatically detect the probability of HCV in multiclass data. The proposed model outperformed XGBoost, which was the second-best performing model, with a difference of 0.02 in terms of accuracy. A recent study employed k-means extreme machine learning for chronic hepatitis diagnosis with an accuracy of 72.36% [19]. However, these studies did not take into consideration the interpretability and explainability of the systems.

A performance evaluation study of various ML models was carried out on a dataset from the Jordan University Hospital [4]. Sequential forward selection and oversampling were applied, with the models including LR, KNN, DT, RF, and neural network all achieving an accuracy of approximately 82.0%. Furthermore, the study presented a global model interpretation by employing SHAP to show the feature importance and impact on the models' decisions.

The motivation for this project is the need for a diagnostic system for HCV detection and its interpretability. The objectives of this project are to collect HCV data, analyze and preprocess the data, design and develop classification models for the detection of hepatitis, and evaluate the models' performances. In addition, the interpretation and explanation of the models' outcomes will be presented using SHAP. Furthermore, a web application for HCV diagnosis was developed using React and FastAPI for the best-performing classification model.

II. Materials and Methods

The following sections describe the dataset collected, analysis and preprocessing steps, followed by the modelling steps for the prediction and diagnosis of hepatitis.

Brief and succinct descriptions of the ML and interpretability techniques are also provided. Data analysis, model development and model interpretation were carried out in Python [20], employing standard frameworks such as Pandas [21], NumPy, Matplotlib [22] and Scikit-learn [23], and SHAP [24] for global feature and local feature importance.

A. Patients Information and Preprocessing

The UCI Machine Learning Repository, a reputable open-access portal that offers benchmark datasets for machine learning research, is where the dataset used in this study was acquired [25, 26, 27]. Researchers specifically gathered the Hepatitis C Virus (HCV) dataset for use in both academic and clinical research. It consists of patient laboratory test results gathered from healthcare facilities to aid in the diagnosis of liver disease, including HCV and associated disorders.

The dataset has 14 attributes and 615 instances in total, where two of the features are categorical data (category and sex) and the remaining twelve features are numerical data. The Patient ID/No column was dropped as it contains a serial number of patients and provides no information relevant to the diagnosis of the patient's condition. The remaining 11 numerical data comprise the age (in years) of the patients and the results of laboratory liver function tests (LFTs) carried out, which include the albumin test (ALB), alkaline phosphatase (ALP), alanine transaminase (ALT), aspartate aminotransferase (AST), bilirubin (BIL), cholinesterase (CHE), cholesterol (CHOL), creatinine (CREA), gamma-glutamyl transferase (GGT), and protein test (PROT). The LFTs are blood tests used to measure different types of enzymes, proteins, and other substances made by the liver [4, 28, 29]. The category feature is the dependent variable having four (4) classes with the following labels: Hepatitis (Hep), Cirrhosis (Cir), Fibrosis (Fib), Blood donor (BD), and Suspect blood donor (SBD).

The dataset was refined to enable proper functioning of the model the data would be used on with the following steps: loading of the dataset, dropping of duplicates, dropping or filling null values, encoding, etc. No duplicate data was found, while there was a total of 26 missing data, which were removed, resulting in a total of 589 instances remaining. This choice was taken after determining that the missing values were not concentrated in specific features, rendering imputation unreliable without creating bias. The percentage of missing data (~4.2%) was low enough that statistical power would not be significantly impacted by deletion. In the context of medical data, imputation could lead to distorted values that may mislead the model, particularly for sensitive biochemical markers.

The sex attributes were encoded as 1 and 0 for the male and female genders, respectively. The descriptive statistical analysis of the cleaned data is shown in Table 1, while the distributions of the features are presented in Figure 1. The age of the patient population ranged from 23 to 77 years, with a mean of 47 years and a spread of years. There are quite a handful of outliers in various features such as ALB, ALP, ALT, AST, BIL, CREA, and GGT, but a choice to keep these records is made as we are dealing with medical data.

Feature engineering is a significant task in optimizing the correctness of a predictive algorithm on a dataset by transforming the feature space. The extraction of features is a very significant step in ML classification for selecting the best features or attributes with the best predictive power, which can be determined using the Pearson correlation coefficient (PCC) [30, 31]. The PCC is a measure of the degree of relationship between two features, X and Y, with values ranging between -1 and +1, and has been calculated as shown in Figure 2 using Equation 1 [32].

$$\rho_{X,Y} = \frac{\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \qquad (1)$$

Table 1. Descriptive Statistics of the Pre-processed Numerical

	Age	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT
Mean	47.42	41.62	68.12	26.58	33.77	11.02	8.20	5.39	81.67	38.20	71.89
Std	9.93	5.76	25.92	20.86	32.87	17.41	2.19	1.13	50.70	54.30	5.35
Min	23	14.90	11.30	0.90	10.60	0.80	1.42	1.43	8.00	4.50	44.80
25%	39	38.80	52.50	16.40	21.50	5.20	6.93	4.62	68.00	15.60	69.30
50%	47	41.90	66.20	22.70	25.70	7.10	8.26	5.31	77.00	22.80	72.10
75%	54	45.10	79.90	31.90	31.70	11.00	9.57	6.08	89.00	37.60	75.20
Max	77	82.20	416.60	325.30	324.00	209.00	16.41	9.67	1079.10	650.90	86.50

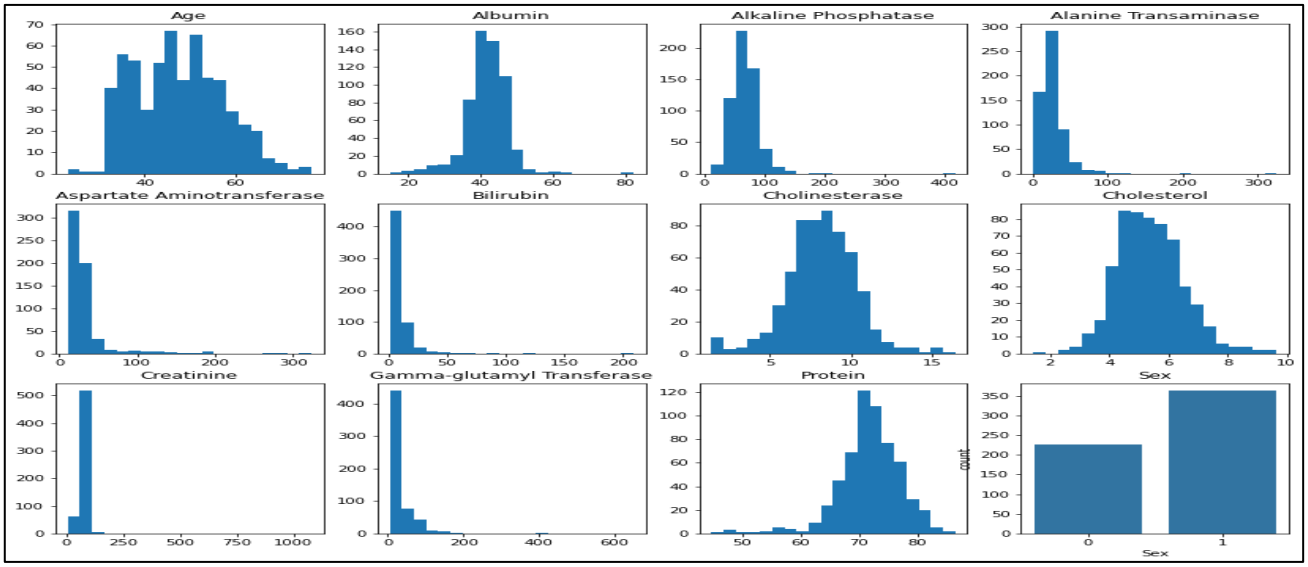

Fig. 1. Frequency Histograms of HCV Dataset Features.

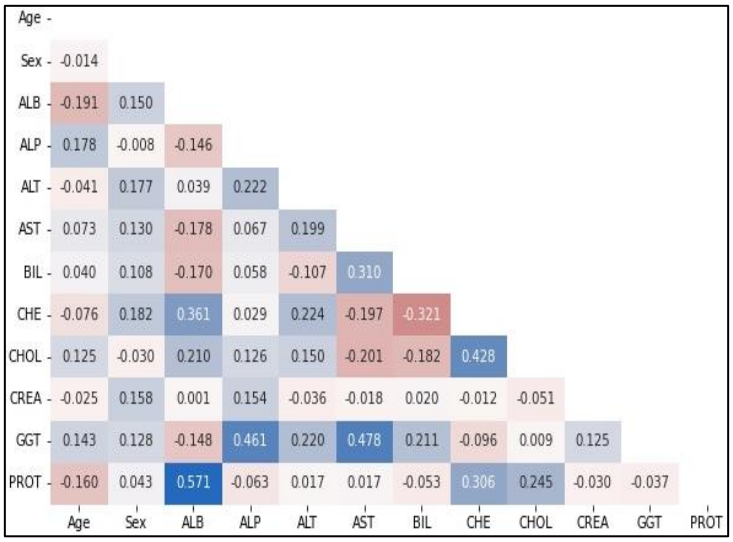

Fig. 2. Bivariate Correlation of the Numerical Features

Where μ_X and μ_Y are the means of X and Y respectively, σ_X and σ_Y are the standard deviations of X and Y respectively and E is the expectation.

Of the 12 variables or features, ALB was moderately correlated with PROT (ρ=0.571) and weakly correlated with CHE (ρ=0.361).

No feature was removed from the dataset as there were no highly correlated features, that is, features with correlation coefficients ranging between $0.7 \leqslant \rho \leqslant 1$.

Figure 3 shows an imbalanced distribution of the categories with the blood donor amounting to 526 (89.3%) instances. The eligible 589 patients' data were randomly split

into a training set (N = 471) and a test set (M = 118). Due to the highly imbalanced nature of the training set, oversampling was done to bring the minority categories to the same amount as the majority category (N_(Blood Donor) = 421), and this procedure is referred to as data augmentation [6]. Synthetic Minority Oversampling Technique (SMOTE) [33] was subsequently adapted to the training set, thereby bringing the total number of training samples to 2105 with all categories having a total number of 421 instances. SMOTE was selected because it effectively creates synthetic samples of the minority class by interpolating between existing samples, avoiding precise duplication and lowering the risk of overfitting. In comparison to random oversampling, SMOTE improves the model's generalization capabilities by maintaining the structure of the feature space and balancing the distribution of classes.

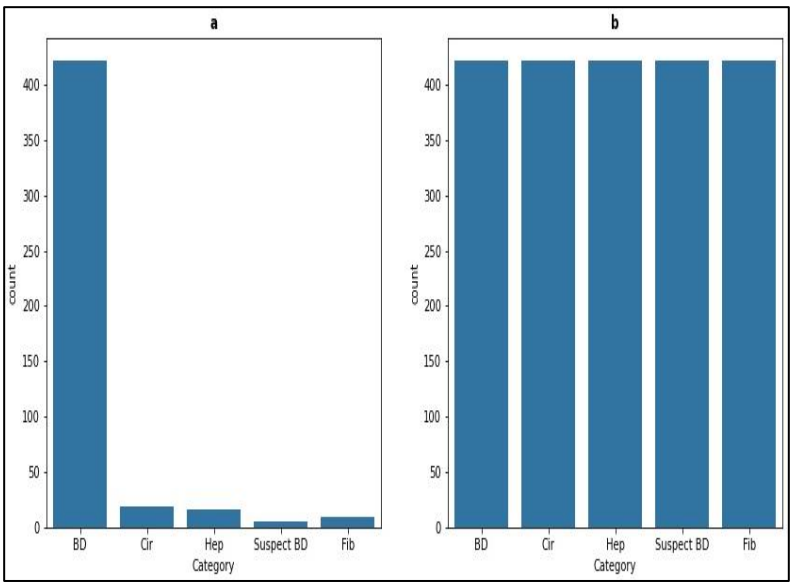

Fig. 3. Distribution of Target Classes in Train Set (a) before SMOTE (b) after SMOTE

The performance of many ML algorithms is improved by scaling the numerical input variable to a standard range, especially for models which utilize the weighted sum of the input or distance measures. The data standardization scales each input variable as stated in Equation 2. This shifts the distribution to be centered around the mean with a standard deviation of 1. The feature scaling (FS) was achieved using the StandardScaler library available in scikit-learn.

$$z = \frac{x - \mu}{\sigma} \quad (2)$$

where z is the standardized value of the original value x, μ and σ are the average and standard deviation of x respectively.

B. Classification Models' Development

ML algorithms are efficient tools for learning patterns from data and predicting future outcomes. There are several applications of ML and AI including classification, regression, clustering and anomaly detection. Classification is a supervised learning method for the prediction of the target label of unseen data from previous data; the target labels can be either binary or multiclass [34, 35].

Seven ML and ANN classification methods were used to build models to classify and predict the HCV status of the patient in this paper. The models employed are the multivariable logistic regression (LR) [36], decision tree classifier (DT) [37], random forest classifier (RF) [38, 39], gradient boosting classifier (GB) [40], k-nearest neighbor (KNN) [35], support vector machine classifier (SVC) [41] and multilayer perceptron (MLP) [42]. They were chosen to represent a broad spectrum of learning paradigms, including:

- Linear models (logistic regression) for baseline interpretability,
- Tree-based models (DT, RF, GB) for handling nonlinear interactions and feature hierarchies,
- Distance-based learners (KNN) for intuitive classification based on feature similarity,
- Margin-based classifiers (SVM) known for robust generalization in high-dimensional spaces,
- Neural networks (MLP) for capturing complex non-linear patterns.

This diverse selection enables a comprehensive performance comparison and facilitates model interpretability and robustness analysis.

The LR model is used for classification by estimating the probability of an event happening and is expressed mathematically using

$$P(y = 1|x_1, x_2, \dots, x_n) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}} \quad (3$$

Where, $x_1, x_2, \dots, x_n$ are the independent variables, and $\beta_1, \beta_2, \dots \beta_n$ are their respective coefficients. It maps any real-valued number into the [0, 1] interval, making it suitable for predicting probabilities. The DT is a tree data structure made up of several nodes and branches.

The decision tree algorithm utilizes a divide and conquer method and repeatedly partitions the input variables to classify or predict the output parameter while RF is an ensemble ML

technique containing numerous decision trees to predict a target variable. The RF uses the Bagging method where homogenous weak learners' models (in this case, decision trees) independently learn from one another in parallel. The final prediction is achieved by a model-averaging approach. The GB tree model is a stacked learning approach where a robust predictive model is formed by combining several individual weak learning trees (DT).

The KNN algorithm uses the distance between data points to make classifications. The class labels are selected based on the majority vote of the selected number of neighbours. The Euclidean distance is used to measure proximity. In SVM, each data point is plotted in n-dimensional space with n as the number of attributes or inputs. The model then performs classification by finding the optimal hyper-plane that differentiates the target labels using the decision rule specified in (4).

$$y = sign(wx + b) \quad (4)$$

Where, y is the predicted class label. sign is the sign function, which returns +1 or -1 depending on the input. The MLP is a type artificial neural network (ANN) commonly used for tasks like classification and regression. An MLP consists of at least three layers: an input layer, one or more hidden layers, and an output layer. The output of a single neuron in an MLP is specified in (5)

$$z = f(\sum_{i=1}^{n} w_i x_i + b) \quad (5)$$

Where z is the output of the neuron. f is the activation function. w_i are the weights associated with the inputs. x_i are the inputs to the neuron. b is the bias term.

The selected models were fitted and trained on the train data set while they were tested and evaluated using the test set. To optimize the performance and correctness of the models, the various algorithms' hyperparameters were tuned and set as depicted in Table 2. Hyperparameters were optimized via GridSearch Cross-Validation, which ensured a systematic and unbiased exploration of the model's parameter space. The GridSearch CV was used to avoid the models memorizing the data, thereby resulting in overfitting. For all computations, the random state was set to 42 to allow for the reproducibility of results.

The settings notably impacted:

- Model convergence (e.g., max_iter=3000 for LR and MLP ensured full convergence),
- Model complexity (e.g., max_depth=11 in DT and RF controlled overfitting),
- Learning dynamics (e.g., learning_rate=0.1 in GB balanced speed and stability),
- Neighbor selection in KNN (neighbors=2) enhanced sensitivity due to small class sizes.

Overall, proper hyperparameter tuning was crucial in attaining high accuracy and stability across all models, especially under imbalanced conditions.

Table 2. Selected Classification Algorithms' Hyperparameters

Model	Hyperparameters	Value
LR	max_iter	3000
	C	1.7
DT	max_depth	11
RF	n_estimators	350
	max_depth	10
GB	max_depth	4
	learning_rate	0.1
	n_estimators	200
KNN	n_neighbors	2
	weights	distance
SVC	C	1.9
	kernel	rbf
MLP	activation	relu
	solver	adam
	learning_rate	constant
	max_iter	200

C. SHAP for Model Interpretability

Explaining and interpreting the decision-making process of a model is oftentimes a difficult process as there is no standard metric for the measurement of explainability and interpretability [14]. Understanding why models make certain decisions is as important as the models' accuracy, especially in the medical field [15]. The interpretability of all the models used in this study was assessed using SHAP based on the Shapley values from game theory [16, 17]. SHAP is used to explain the outcome of ML models. It connects optimal credit allocation with local explanations and assigns each input or attribute an importance value for a particular prediction.

The KernelExplainer method in SHAP which is a flexible method capable of working with all types of models is used to determine the individual feature contributions using Shapley values. The KernelExplainer builds a weighted linear regression to compute the importance of each feature. Due to how computationally expensive the process is, a sample of 500 training sets was used as background data for the SHAP model inference. This was implemented on the models without feature scaling. The SHAP algorithm is presented as:

Algorithm 1: SHAP Algorithm

1. **Model Prediction**
   Obtain the prediction $\hat{y}$ for the instance to be explained
2. **Initialize Parameters**
   Let F be the set of all features.
   For each feature x_i, initialize its Shapley value Θ_i to 0.
3. **Calculate the Shapley Value**
   For each feature x_i:
   a. Consider all possible subsets S of features excluding x_i
   b. For each subset S:
      i. Calculate the model prediction using only the features in S:f(S)
      ii. Calculate the model prediction using the features in S plus feature x_i: $f(SU\{i\})$.
      iii. Compute the difference in predictions: $\Delta = f(SU\{i\} - f(S)$.

iv. Weight the difference by the combinatorial factor $\frac{|S|!(|F|-|S|-1)!}{|F|!}$

v. Accumulate the weighted difference to the Shapley value of feature x_i:

$$\Theta_i += \frac{|S|!\,(|F|-|S|-1)!}{|F|!}\Delta$$

4. Normalize Shapley Values

For each feature x_i, average its accumulated Shapley value over all subsets S to get the final Shapley value

$$\Theta_i = \sum_{S \sqsubseteq F \setminus \{i\}} \frac{|S|!\,(|F|-|S|-1)!}{|F|!}\Delta$$

III. Result and Discussion

A. Model Evaluation

In Table 3, the confusion matrices for the individual models with and without feature scaling are presented. The DT, KNN, MLP and SVC had a slightly better prediction accuracy when feature scaling was incorporated compared to the performance without feature scaling. The mean performance of the classification models was determined by metrics including accuracy, sensitivity and specificity as presented in Table 4 and Figure 4. All models performed better than the baseline accuracy (accuracy gotten by always predicting the majority class which is the blood donor group) of 0.893. For classification without applying feature scaling, GB narrowly performed the best with an accuracy of 0.97, followed by SVC, MLP, RF and KNN; all with an accuracy of 0.95, LR (0.94) and DT (0.91). The results obtained for training while applying feature scaling were similar except for MLP and DT improving their accuracy performance to 0.97 and 0.92 respectively.

In terms of the model's ability to correctly detect cases of patients that have an HCV-related disease or suspected donor, referred to as the sensitivity, SVC (with and without FS) and MLP (with FS) performed best with a sensitivity of 1.00, meaning that all patients with HCV were detected.

Table 3. Confusion Matrices for the Models

Methods	Actual Class	Predicted Class without FS					Predicted Class with FS				
		BD	CIR	FIB	HEP	SBD	BD	CIR	FIB	HEP	SBD
LR	BD	104	0	0	1	0	104	1	0	0	0
	CIR	1	4	0	0	0	1	4	0	0	0
	FIB	0	0	1	2	0	1	0	0	2	0
	HEP	0	1	2	1	0	0	0	2	2	0
	SBD	0	0	0	0	1	0	0	0	0	1
DT	BD	101	2	1	1	0	101	2	0	2	0
	CIR	1	4	0	0	0	1	4	0	0	0
	FIB	0	0	1	2	0	0	1	1	1	0
	HEP	2	0	1	1	0	1	0	2	1	0
	SBD	0	1	0	0	0	0	0	0	0	1
RF	BD	105	0	0	0	0	105	0	0	0	0
	CIR	1	4	0	0	0	1	4	0	0	0
	FIB	0	0	1	2	0	0	0	1	2	0
	HEP	2	0	1	1	0	2	0	1	1	0
	SBD	0	0	0	0	1	0	0	0	0	1
GB	BD	105	0	0	0	0	105	0	0	0	0
	CIR	0	5	0	0	0	0	4	1	0	0
	FIB	0	0	2	1	0	0	0	3	0	0
	HEP	1	1	1	1	0	1	1	1	1	0
	SBD	0	0	0	0	1	0	0	0	0	1
KNN	BD	105	0	0	0	0	105	0	0	0	0
	CIR	1	4	0	0	0	1	4	0	0	0
	FIB	0	0	2	1	0	1	0	0	2	0
	HEP	1	1	1	1	0	0	1	1	2	0
	SBD	1	0	0	0	0	0	0	0	0	1
SVC	BD	104	0	0	1	0	105	0	0	0	0
	CIR	0	4	0	1	0	0	4	1	0	0
	FIB	0	0	2	1	0	0	0	0	3	0
	HEP	0	1	1	2	0	0	1	1	2	0
	SBD	0	1	0	0	0	0	0	0	0	1
MLP	BD	103	0	0	0	2	105	0	0	0	0
	CIR	0	5	0	0	0	0	4	1	0	0
	FIB	0	0	2	1	0	0	0	2	1	0
	HEP	1	1	1	1	0	0	1	1	2	0
	SBD	0	0	0	0	1	0	0	0	0	1

Table 4: Performance Evaluation

	Without Feature Scaling			**With Feature Scaling**		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
LR	0.94	0.92	0.99	0.94	0.84	0.99
DT	0.91	0.77	0.96	0.92	0.84	0.96
RF	0.95	0.77	1.00	0.95	0.77	1.00
GB	0.97	0.92	1.00	0.97	0.92	1.00
KNN	0.95	0.77	1.00	0.95	0.84	1.00
SVC	0.95	1.00	0.99	0.95	1.00	1.00
MLP	0.95	0.92	0.98	0.97	1.00	1.00

Contrary to sensitivity, specificity measures the classifier's ability to detect patients without a condition. Three models (RF, GB and Figure KNN) with and without feature scaling, as well as scaled SVC and MLP had the best specificity having achieved a value of 1.00 (all 105 blood donors without any traces of HCV infection were correctly identified). These two terms are very important for consideration in medical diagnostics. Considering all the metrics and the confusion matrices, the MLP (with FS) has the best performance, achieving an accuracy of 0.97, and sensitivity and specificity of 1.00.

Using the same or comparable HCV datasets, earlier research has reported accuracy rates between 85.0% and 95.0%; however, most of these studies either did not address class imbalance or used interpretability methodologies that were too restrictive [43, 44, 45, 46]. This work stands out for the following reasons

- Using SMOTE to address imbalanced data improves sensitivity
- Evaluating seven different machine learning models under both scaled and unscaled conditions.
- In contrast to previous studies, SHAP is used for both local and global interpretability.
- Deploying a real-time diagnostic web app, making the solution immediately usable by clinicians.

In comparison to relevant literature, the study's top-performing model obtained 97.00% accuracy, 100.00% sensitivity, and 100.00% specificity, all of which are on the upper bound.

B. Model Interpretability using Shapley's Values

With the above evaluation results presented, the big questions remain unanswered. Questions such as: (i) How do the models make the decisions? (ii) What features or attributes are responsible for the decision? (iii) How much importance or impact do individual features have? (iv) Can the results be trusted? SHAP can answer these questions using both global and local model interpretability which is based on the addition or aggregation of Shapley values

C. Global Model Interpretability

The average behaviors of the models were explained using the global model interpretability function. SHAP not only shows the feature's importance but takes it a step further showing the feature's influence whether positive or negative. Figure 5 depicts the feature importance of the models for all the category classes (BD, Hep, Cir, Fib and Suspect BD). The various classes are color-coded according to the legend. The models found the Albumin test, Alkaline Phosphatase, Alanine Transaminase and Aspartate Aminotransferase as the most important features with the most influence on the outcome of the predictions with a few variations or exceptions.

Analyzing the feature importance of each model deeper reveals that the 5 most influential features (ALB, AST, ALP, BIL and CHE) of the DT and RF were similar, and since the RF is a stack of DTs, the feature influence would be similar. GB, however, surprisingly found PROT as the most important closely followed by AST and CHE. The SVC and MLP had similar features (AST, ALT, ALP, GGT and BIL) with the most influence on the predictions.

In general, all models found sex and cholesterol as the two least important or influential features for the prediction of HCV. This goes to show that the hepatitis C virus is not a respecter of gender. KNN and SVC models hardly utilize these features (sex and CHOL) in addition to Cholinesterase at all for prediction, while sex was the least important for all models except MLP where the least was CHOL.

Figure 6 depicts the influence of individual values on the outcome where each point represents an instance of each feature. The blue points represent low values of the feature while the red points represent high values as shown by the legend. The x-axis shows the value of the influence of each instance on the predictions. Generally, the lower values of AST and GGT have a positive influence while the higher values have a negative impact on the outcomes for all the seven models considered. This contrasts with the ALP, ALB and PROT, in which the higher values have a positive influence while the lower values are responsible for a negative impact on the model output. In addition, Sex and CHOL are clustered around 0.0 (i.e. they have low SHAP values). SHAP values for all models thereby verifying that the two features have minimal impact on the model's outputs. For clinical acceptance, SHAP values offer clear, feature-level insights into the model's decision-making process. From a medical perspective. Standard liver function tests, including albumin (ALB), ALP, ALT, AST, and BIL, were found to be the most influential markers. When screening for or diagnosing liver-related diseases, clinicians can give priority

to these markers. The negligible influence of sex and cholesterol aligns with known HCV pathology, suggesting clinicians can focus less on these factors in isolation. Local SHAP values enable individualized patient-level explanations, which are essential for precision medicine, allowing physicians to understand why a specific diagnosis was made for a particular patient, thereby improving trust and clinical usability.

D. Local Model Interpretability

SHAP values can be calculated for each prediction to know how the features contribute to that single prediction. All models were also evaluated using the local interpretability method. This is of great advantage and significance in understanding individual predictions. Figure 7 shows the feature impacts for an individual prediction with a base value (average prediction) of $E[f(x)] \approx 0.17$ and the model prediction probability value $f(x) \approx 0$. The plots show the individual feature contributions to arriving at the model prediction probability.

The feature impact of all the features in achieving the prediction probability for the LR is depicted in Figure 7(a). ALP has the most influence on the model's prediction with a Shapley value of +0.24 followed by GGT, ALT and CHE with values-0.15, -0.09 and -0.06 respectively. These are the most predictive features for the individual instance. CHOL, Sex and CREA have little to no impact on the outcome of the LR model. These features all have a Shapley value of ±0 and thus are the least predictive features. Interestingly, these three least impactful features are consistent with the same values across all seven models implemented. It can be concluded that the values of CHOL (5), sex (1 - Male) and CREA (74) do not have an impact on the final prediction

E. Web-Based Hepatitis C Virus Diagnosis Assistant

A web-based hepatitis C virus diagnosis assistant was built based on the best-performing model, the multilayer perceptron with feature scaling to enable clinicians to predict the risk of Hepatitis C in patients using react for the frontend and fastAPI was used for the backend. Vercel was used for the hosting of the frontend while render was used for the deploying of the fastAPI. The diagnosis assistant has been designed to be user-friendly, interactive and efficient for end users including physicians and laboratory scientists. The Hepatitis C virus diagnosis assistant is available at https://hcv-prediction-ui-y8yy.vercel.app/. Figure 8 depicts the test page and the result page.

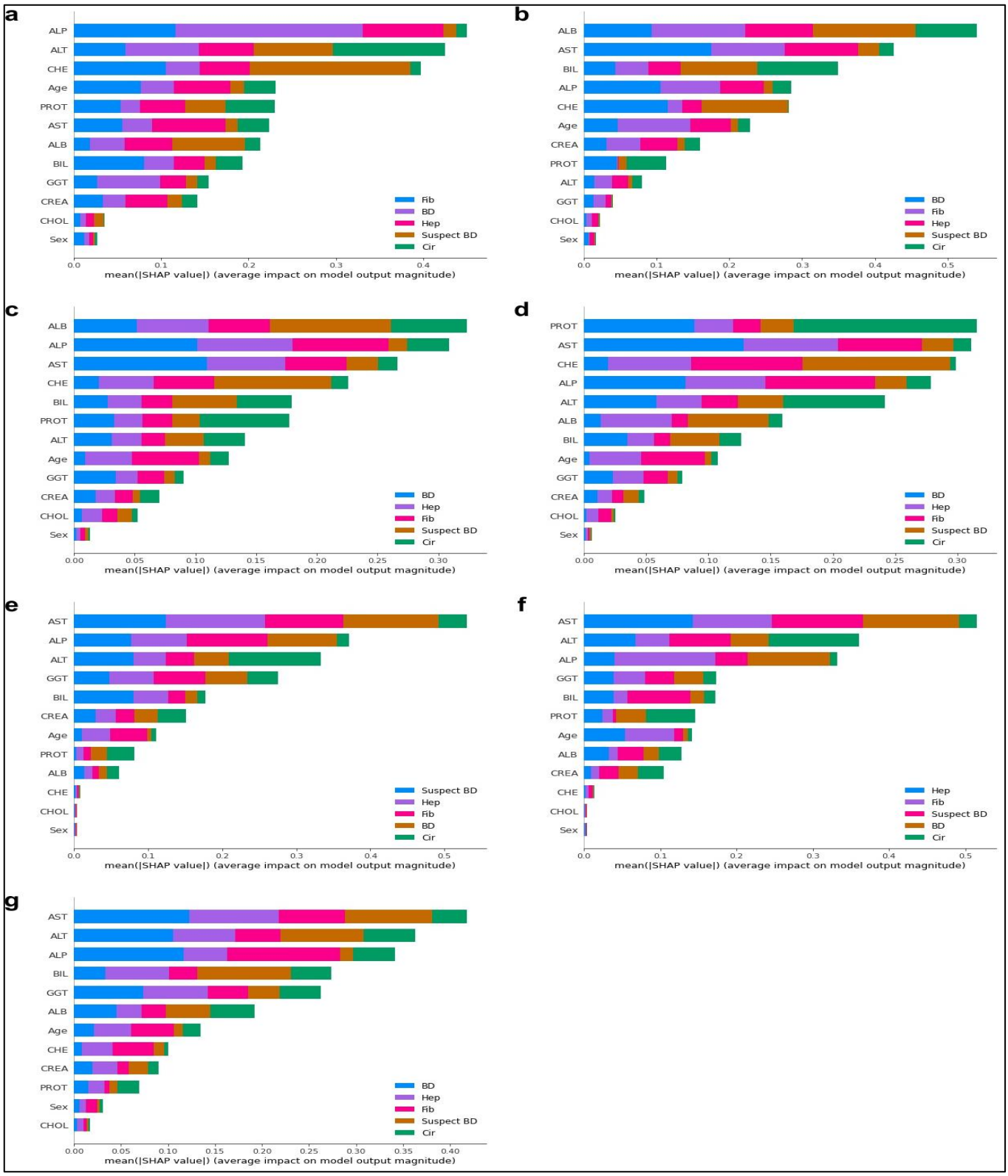

Fig. 4. Global Feature Importance using SHAP (Shapley Additive Explanations) for the following Models without Feature Scaling (a) LR (b) DT (c) RF (d) GB (e) KNN (f) SVC (g) MLP

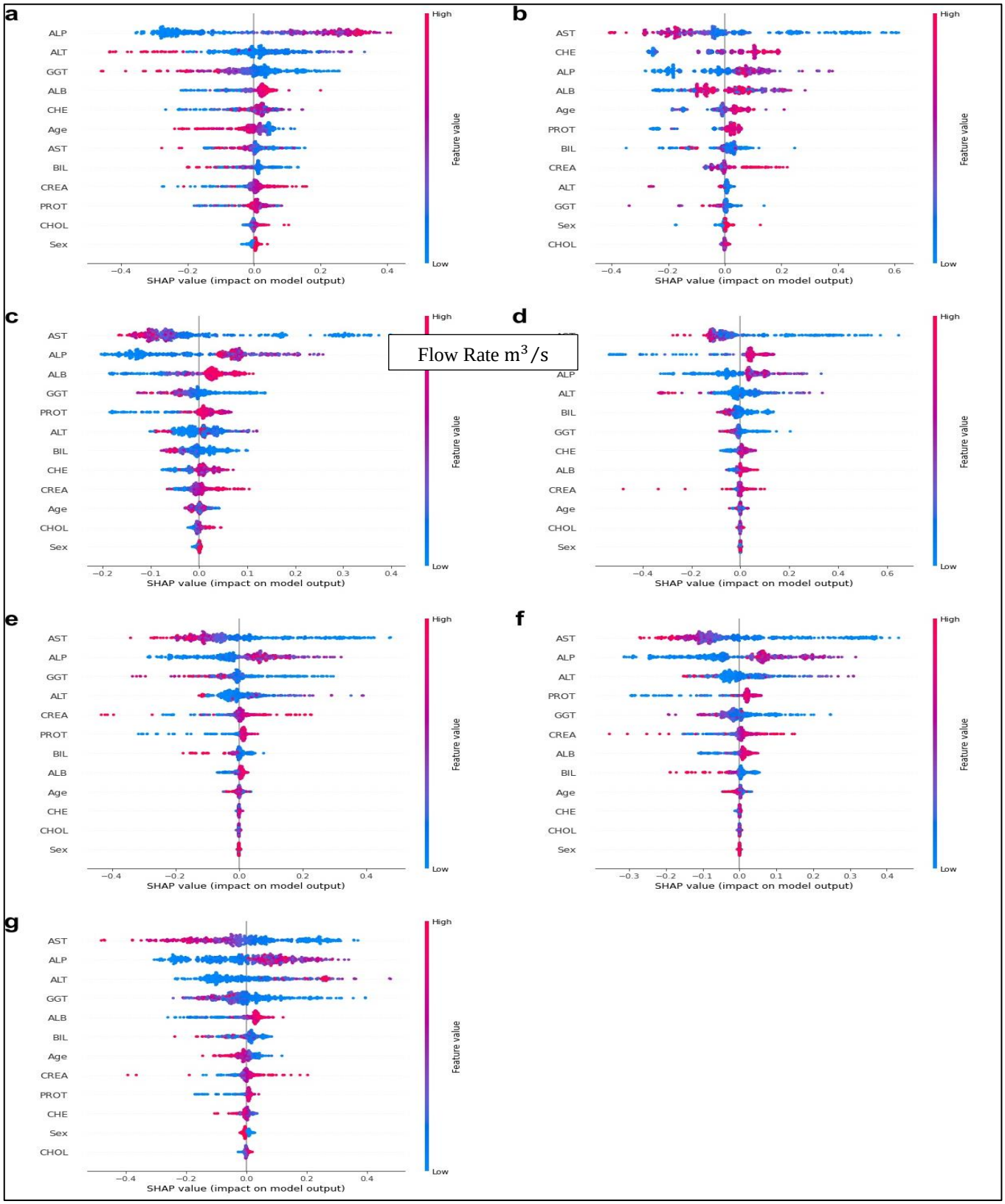

Fig. 5. Summary Plots using SHAP (Shapley Additive Explanations)
for the following Models without Feature Scaling (a) LR (b) DT (c) RF (d) GB (e) KNN (f) SVC (g) MLP

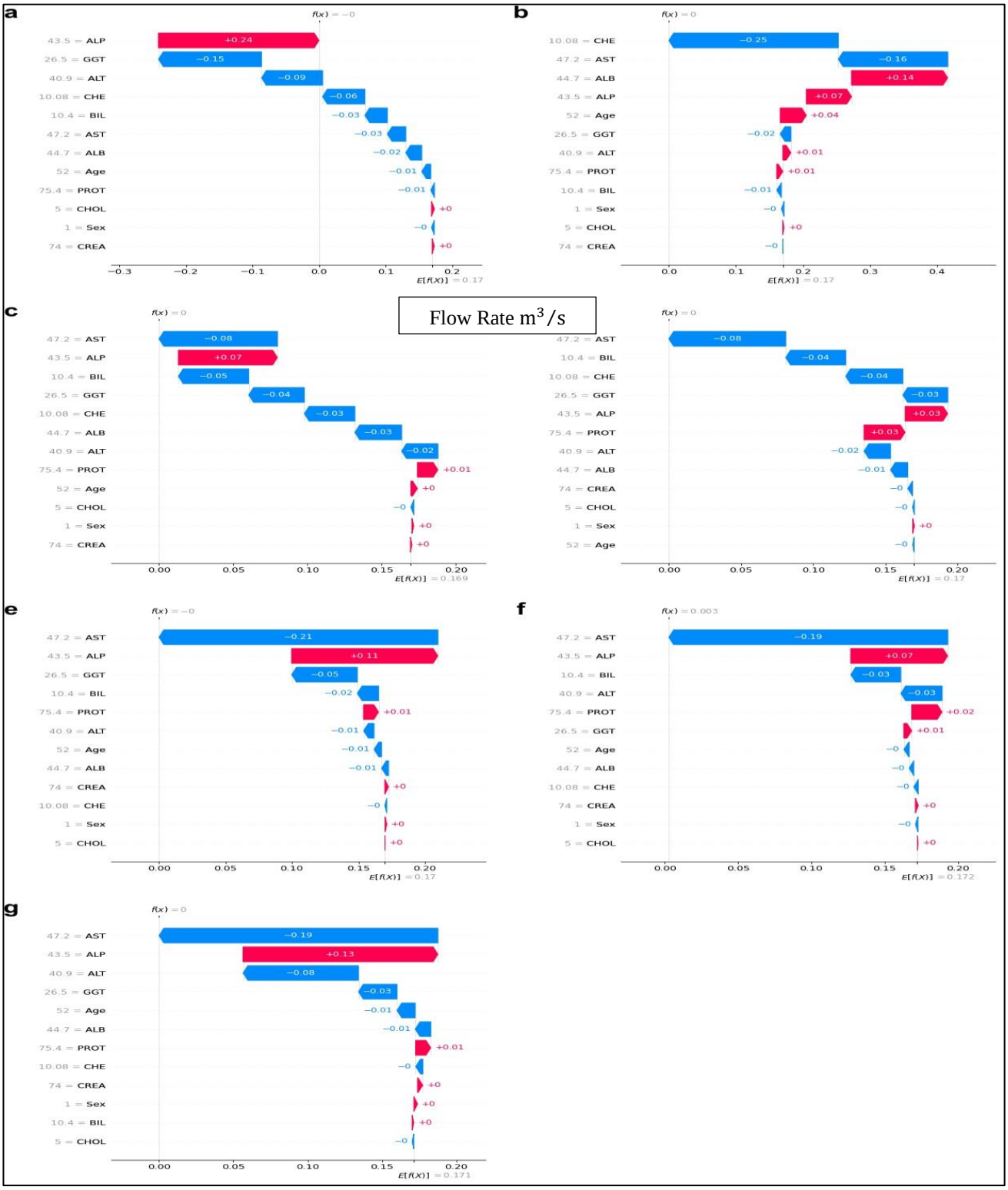

Fig. 6. Feature Importance for a Single Test Instance
for the following Models without Feature Scaling (a) LR (b) DT (c) RF (d) GB (e) KNN (f) SVC (g) MLP

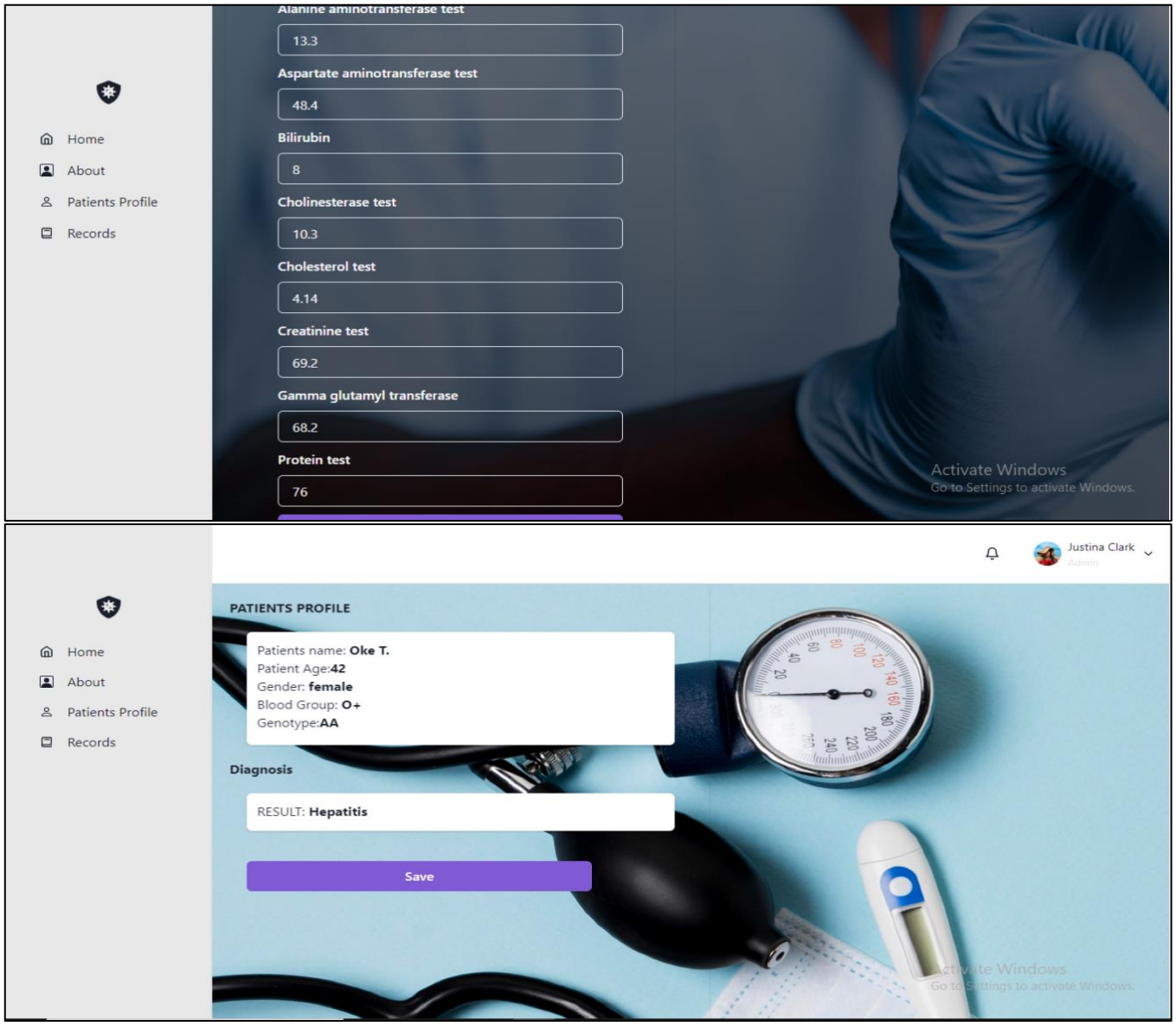

Fig. 7. HCV Diagnosis Assistant - (Top) Test Page (Bottom) Result Page

IV. Conclusion

Early detection and diagnosis of diseases and health conditions result in a significant reduction in the mortality rate among patients. The process of detection and diagnosis of hepatitis C infection can be improved by adapting the use of artificial intelligence and machine learning. Artificial intelligence algorithms are efficient tools for the early diagnosis of health conditions. This study conducted a comparative analysis of AI methods for the detection of HCV using accuracy, sensitivity and specificity as metrics. The study showed that the multilayer perceptron algorithm implemented with feature scaling had the best level of the performance metrics used at the same time.

Of great importance to the end-users is the understanding of how the models are making their decisions and what the feature contributions are. SHAP based on Shapley values was used for global and local model interpretability. The models found the albumin test, alkaline phosphatase, alanine.

Transaminase and Aspartate Aminotransferase are the most important features with the most influence while sex and cholesterol are the two least important features for the prediction of HCV.

Despite promising results, the study admits a number of shortcomings. Larger datasets are required to enhance external validity, even though this size is enough for model training. Analysis of population subgroups was limited because only age and gender were available.

In the future, we aim to explore improving model generalizability, larger and more varied datasets covering various ethnicities, comorbidities, and geographic populations. Additional clinical features, such as patient history, symptoms, or treatment data, will also be incorporated. Deploying the web-based assistant in real-time and clinically testing it in hospital settings for validation and feedback is also another direction.

REFERENCES

[1] World Health Organization, "Hepatitis C," accessed Apr. 20, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/hepatitis-c

[2] H. Mamdouh Farghaly, M. Y. Shams, and T. Abd El-Hafeez, "Hepatitis C Virus prediction based on machine learning framework: a real-world case study in Egypt," *Knowledge and Information Systems*, 2023, doi: 10.1007/s10115-023-01851-4.

[3] M. Cedolin, M. E. Genevois, and Z. Canbulat, "Hepatitis C Diagnosis Using Computational Intelligence Techniques," 2024, pp. 29–36, doi: 10.1007/978-3-031-67192-0_4.

[4] 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.

[5] S. Agrawal, R. K. Dhiman, and J. K. Limdi, "Evaluation of abnormal liver function tests," *Postgraduate Medical Journal*, vol. 92, no. 1086, pp. 223–234, 2016, doi: 10.1136/postgradmedj-2015-133715.

[6] L. Chen, P. Ji, and Y. Ma, "Machine Learning Model for Hepatitis C Diagnosis Customized to Each Patient," *IEEE Access*, vol. 10, pp. 106655–106672, 2022, doi: 10.1109/ACCESS.2022.3210347.

[7] A. O. Oyedeji, M. O. Osifeko, O. Folorunsho, O. R. Abolade, and O. O. Ade-Ikuesan, "Design and Implementation of a Medical Diagnostic Expert System," *Journal of Engineering Science*, vol. 10, no. 2, pp. 103–109, 2019.

[8] M. A. Myszczynska et al., "Applications of machine learning to diagnosis and treatment of neurodegenerative diseases," *Nature Reviews Neurology*, vol. 16, no. 8, pp. 440–456, 2020, doi: 10.1038/s41582-020-0377-8.

[9] S. K. Mun, K. H. Wong, S. C. B. Lo, Y. Li, and S. Bayarsaikhan, "Artificial Intelligence for the Future Radiology Diagnostic Service," *Frontiers in Molecular Biosciences*, vol. 7, 2021, doi: 10.3389/fmolb.2020.614258.

[10] S. Ekins et al., "Exploiting machine learning for end-to-end drug discovery and development," *Nature Materials*, vol. 18, no. 5, pp. 435–441, 2019, doi: 10.1038/s41563-019-0338-z.

[11] A. J. P. L et al., "Medical information retrieval systems for e-Health care records using fuzzy based machine learning model," *Microprocessors and Microsystems*, 2020, doi: 10.1016/j.micpro.2020.103344.

[12] H. Haga et al., "A machine learning-based treatment prediction model using whole genome variants of hepatitis C virus," *PLoS ONE*, vol. 15, no. 11 November, 2020, doi: 10.1371/journal.pone.0242028.

[13] M. O. Edeh et al., "Artificial Intelligence-Based Ensemble Learning Model for Prediction of Hepatitis C Disease," *Frontiers in Public Health*, vol. 10, 2022, doi: 10.3389/fpubh.2022.892371.

[14] C. E. Charlton, M. T. C. Poon, P. M. Brennan, and J. D. Fleuriot, "Development of prediction models for one-year brain tumour survival using machine learning: a comparison of accuracy and interpretability," *Computer Methods and Programs in Biomedicine*, p. 107482, 2023, doi: 10.1016/j.cmpb.2023.107482.

[15] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *31st Conference on Neural Information Processing Systems (NIPS 2017)*, vol. 32, no. 2, pp. 1208–1217, 2017.

[16] M. Sundararajan and A. Najmi, "The many shapley values for model explanation," *37th International Conference on Machine Learning, ICML 2020*, vol. PartF16814, pp. 9210–9220, 2020.

[17] T. R. Noviandy et al., "Interpretable machine learning approach to predict Hepatitis C virus NS5B inhibitor activity using voting-based LightGBM and SHAP," *Intelligent Systems with Applications*, vol. 25, p. 200481, Mar. 2025, doi: 10.1016/j.iswa.2025.200481.

[18] T. H. S. 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. 91045–91058, 2022, doi: 10.1109/ACCESS.2022.3202295.

[19] J. Cai, T. Chen, Y. Qi, S. Liu, and R. Chen, "Fibrosis and inflammatory activity diagnosis of chronic hepatitis C based on extreme learning machine," *Scientific Reports*, vol. 15, no. 1, p. 11, Jan. 2025, doi: 10.1038/s41598-024-84695-4.

[20] S. Raschka, J. Patterson, and C. Nolet, "Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence," *Information (Switzerland)*, vol. 11, no. 4, 2020, doi: 10.3390/info11040193.

[21] J. Reback *et al.*, "pandas-dev/pandas: Pandas 1.0.5," *Zenodo*, Jun. 2020, doi: 10.5281/ZENODO.3898987.

[22] J. Ranjani, A. Sheela, and K. P. Meena, "Combination of NumPy, SciPy and Matplotlib/Pylab—A good alternative methodology to MATLAB—A Comparative analysis," *Proc. 1st Int. Conf. Innovations Inf. Commun. Technol. (ICIICT)*, 2019, doi: 10.1109/ICIICT1.2019.8741475.

[23] G. Varoquaux, L. Buitinck, G. Louppe, O. Grisel, F. Pedregosa, and A. Mueller, "Scikit-learn," *GetMobile: Mobile Comput. Commun.*, vol. 19, no. 1, pp. 29–33, 2015, doi: 10.1145/2786984.2786995.

[24] S. Mangalathu, S. H. Hwang, and J. S. Jeon, "Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach," *Eng. Struct.*, vol. 219, 2020, doi: 10.1016/j.engstruct.2020.110927.

[25] R. Lichtinghagen and G. Hoffmann, "HCV data Data Set," *Machine Learning Repository, University of California, Irvine*, Accessed: Dec. 13, 2022. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/HCV+data

[26] D. Dua and C. Graff, "UCI Machine Learning Repository," [Online]. Available: http://archive.ics.uci.edu/ml

[27] G. Hoffmann, A. Bietenbeck, R. Lichtinghagen, and F. Klawonn, "Using machine learning techniques to generate laboratory diagnostic pathways—a case study," *J. Lab. Precis. Med.*, vol. 3, p. 58, 2018, doi: 10.21037/jlpm.2018.06.01.

[28] B. R. Thapa and A. Walia, "Liver Function Tests and their Interpretation," *Indian J. Pediatr.*, vol. 74, pp. 67–75, 2007.

[29] Hepatitis NSW, "Hepatitis factsheet: Liver Function Tests," 2017.

[30] S. M. Abdelmoneim, M. Kayed, and S. A. Taie, "A Comparative study for Feature Extraction and Classification of Images," *Proc. ACCS/PEIT 2019 - 6th Int. Conf. Adv. Control Circuits Syst. and 5th Int. Conf. New Paradigms Electron. Inf. Technol.*, pp. 105–110, 2019, doi: 10.1109/ACCS-PEIT48329.2019.9062848.

[31] I. Guyon and A. Elisseeff, "An introduction to feature extraction," *Stud. Fuzziness Soft Comput.*, vol. 207, pp. 1–25, 2006, doi: 10.1007/978-3-540-35488-8_1.

[32] P. Schober and L. A. Schwarte, "Correlation coefficients: Appropriate use and interpretation," *Anesth. Analg.*, vol. 126, no. 5, pp. 1763–1768, 2018, doi: 10.1213/ANE.0000000000002864.

[33] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[34] A. Oyedeji, A. Salami, O. Folorunsho, and O. Abolade, "Analysis and Prediction of Student Academic Performance Using Machine Learning," *J. Inf. Technol. Comput. Eng.*, vol. 4, no. 1, pp. 10–15, 2020, doi: 10.25077/jitce.4.01.10-15.2020.

[35] A. Olayiwola, A. Oyedeji, O. Omoyeni, O. Ayemimowa, and M. Olaoluwa, "Comparative Analysis of Machine Learning Models for Detection of Fake News: A Case Study of Covid-19," *J. Inf. Technol. Comput. Eng.*, vol. 7, no. 1, pp. 29–33, 2023, doi: 10.25077/jitce.7.01.29-33.2023.

[36] L. Buitinck *et al.*, "API design for machine learning software: experiences from the scikit-learn project," 2013. [Unpublished/Preprint if applicable].

[37] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 1, pp. 20–28, 2021, doi: 10.38094/jastt20165.

[38] A. Chaudhary, S. Kolhe, and R. Kamal, "An improved random forest classifier for multi-class classification," *Inf. Process. Agric.*, vol. 3, no. 4, pp. 215–222, 2016, doi: 10.1016/j.inpa.2016.08.002.

[39] Z. Masetic and A. Subasi, "Congestive heart failure detection using random forest classifier," *Comput. Methods Programs Biomed.*, vol. 130, pp. 54–64, 2016, doi: 10.1016/j.cmpb.2016.03.020.

[40] U. Singh, M. K. Gourisaria, and B. K. Mishra, "A Dual Dataset approach for the diagnosis of Hepatitis C Virus using Machine Learning," *Proc. 2022 IEEE Int. Conf. Electron. Comput. Commun. Technol. (CONECCT)*, 2022, doi: 10.1109/CONECCT55679.2022.9865758.

[41] A. Tharwat, "Parameter investigation of support vector machine classifier with kernel functions," *Knowl. Inf. Syst.*, vol. 61, no. 3, pp. 1269–1302, 2019, doi: 10.1007/s10115-019-01335-4.

[42] H. Taud and J. F. Mas, "Multilayer Perceptron (MLP)," in *Geomatic Approaches for Modeling Land Change Scenarios*, pp. 451–455, 2018, doi: 10.1007/978-3-319-60801-3_27.

[43] D. B. Smith, B. Donald, J. Bukh, C. Kuiken, A. S. Muerhoff, C. M. Rice, J. T. Stapleton, and P. Simmonds, "Expanded classification of hepatitis C virus into 7 genotypes and 67 subtypes: updated criteria and genotype assignment web resource," *Hepatology*, vol. 59, no. 1, pp. 318–327, 2014, doi: 10.1002/hep.26744.

[44] R. Safdari, A. Deghatipour, M. Gholamzadeh, and K. Maghooli, "Applying data mining techniques to classify patients with suspected hepatitis C virus infection," *Intell. Med.*, vol. 2, no. 4, pp. 193–198, 2022, doi: 10.1016/j.imed.2021.12.003.

[45] O. O. Oladimeji, A. Oladimeji, and O. Olayanju, "Machine Learning Models for Diagnostic Classification of Hepatitis C Tests," *Front. Health Inform.*, vol. 10, no. 1, p. 70, 2021, doi: 10.30699/fhi.v10i1.274.

[46] A. Alotaibi, L. Alnajrani, N. Alsheikh, A. Alanazy, S. Alshammasi, M. Almusairii, S. Alrassan, and A. Alansari, "Explainable ensemble-based machine learning models for detecting the presence of cirrhosis in Hepatitis C patients," *Computation*, vol. 11, no. 6, p. 104, 2023, doi: 10.3390/computation11060104.