Department of Electrical and Computer Engineering

North South University

CSE445 Report

Hepatitis C Prediction Using Machine Learning

Name	ID
Md. Saikot Hossain Sojib	2014055642
Md. Tasin Hossain Toha	2011664042

Faculty:

DR. RIASAT KHAN [RTK]

Associate Professor

ECE Department

Spring 2025

Section	Contributing Member Name	
IEEE Word/LaTeX formatting	Sojib, Toha	
Grammarly check	Toha	Grammarly Score: 96
Abstract	Sojib	
Keywords	Sojib, Toha	
Introduction Motivation	Sojib, Toha	
Paper Review 1	Sojib	[1]
Paper Review 2	Toha	[2]
Introduction Second-Last Paragraph (describe your work)	Sojib, Toha	
Proposed System (Dataset and Preprocessing)	Sojib, Toha	
Proposed System (Model description)	Sojib	Stacked Model, DT, XGBoost
	Toha	Random Forest, SVM, KNN
Results and Discussion	Sojib, Toha	
Figure and Table Title Formatting	Sojib, Toha	
Conclusions	Sojib, Toha	

Hepatitis C Prediction Using Machine Learning

Md. Saikot Hossain Sojib
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
saikot.sojib@northsouth.edu

Md. Tasin Hossain Toha
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
tasin.toha@northsouth.edu

Abstract - Hepatitis C Virus (HCV) is a viral infection related to the liver and can develop into life-threatening situations if not diagnosed in its early stages. Timely diagnosis and proper staging are essential for effective treatment and survival of the patient. This study develops a machine learning prediction model to enhance existing works' limitations, including low-accuracy prediction, data imbalance, and limited interpretability. Our proposed system includes several machine learning techniques: Random Forest (achieved 67.11% accuracy), XG Boost (62.6% accuracy), Support Vector Machine (57.2% accuracy), Decision Tree (achieved 67.11% accuracy and K-Nearest Neighbors (45.94% accuracy). To improve the predictive model, we develop a stacked ensemble model. Random Forest, XG Boost, KNN, and SVM are combined as base learners with XG Boost as the meta-learner, achieving a model with 88% accuracy and an F1 score of 0.88 in model training. The HCV data set was taken from the UCI Machine Learning Repository. It comprises 1385 samples and 29 features, such as clinical and laboratory characteristics (ALT, AST, RNA levels, platelet count). The original dataset consists of four classes for different stages of liver fibrosis, which were transformed into a binary classification to enhance the model training and predictive capacity. For addressing class imbalance, the Borderline SMOTE sampling technique was used to make the data classes balanced for efficient learning. Besides, feature selection was implemented through Recursive Feature Elimination (RFE) with Random Forest to ensure that major features like Baseline Histological Grading and Baseline Histological Staging were not discarded, taking altogether 18 best features. Scaling was also performed with the Min-Max Scaler, and feature ranges were normalized. Local Interpretable Model-Agnostic Explanations (LIME) were used to aid model transparency, which gave the exact local explanations for individual predictions. The experimental results show that integrating feature selection, balanced learning, ensemble learning, and interpretability measures for the task of Hepatitis C staging markedly improves the diagnostic performance compared with traditional single model approaches. This study highlights the benefit of ensemble machine learning in combination with explainable AI methods.

Keywords— Machine Learning, Hepatitis C, Ensemble Learning, Feature Selection, RFE, Borderline SMOTE, LIME

I. Introduction

Hepatitis C virus (HCV) remains a serious health problem worldwide [1], as it frequently results in chronic disease that can progress to severe liver diseases, including cirrhosis and liver cancer. Unlike HAV and HBV, the majority of individuals infected with HCV (~75%) fail to spontaneously resolve the infection and progress to chronic infection [2]. While the burden of HCV is significant, there are currently sparse data on its transmission, pathogenesis, and burden. In addition, the existing diagnostic approaches are expensive and invasive, and early detection is challenged in low-income areas. This raises an increasing interest in leveraging clinical data and machine learning methods for early prediction and staging of the disease.

Lilhore et al. [3] proposed a hybrid machine learning model that combines improved random forest and support vector machine to enhance hepatitis C classification. They used the UCI HCV dataset and implemented SMOTE to handle class imbalance and a ranker method for feature selection. The suggested hybrid model achieved 96.82% accuracy, which is higher than other ML approaches and is very promising for the early and accurate detection of HCV.

Kim et al. [4] developed a prediction model for HBV/HCV among patients with diabetes using machine learning, which was relatively more biased toward early detection in high-risk groups. They used NHANES data (2013–2018), using 1,396 diabetic cases (64 positive, 1,332 negative), and utilized SMOTE to tackle the class imbalance in their data. Four machine learning models were tested: Random Forest, SVM, XG Boost, and LASSO, with a stacked ensemble, using hyperparameter tuning optimization for all the models. LASSO, with an AUC-ROC of

0.810, was consistently top, and illicit drug use, poverty, and race were found to be the most influential contributors. While the findings demonstrate the potential for machine learning to enhance hepatitis screening, the ensemble model did not outperform LASSO.

This project aims to predict Hepatitis C from clinical data in the UCI repository. The Necessary steps were missing value treatment, feature selection using RFE, and class balancing with Borderline SMOTE. Random Forest, followed by XG Boost, SVM, and KNN voter with XG Boost as the meta-learner, a placeholder with high accuracy. Model interpretation was achieved using LIME.

II. Proposed System

The proposed system is a stack ensemble model where RF, SVM, KNN, and XG Boost are base learners, and XG Boost is a meta classifier. It takes advantage of the power of the different models to enhance the HCV stage prediction capabilities and general performance.

A. Dataset

The project works with the Hepatitis C Virus (HCV) dataset [5], which is available from the UCI repository and consists of 1385 samples of patient records, including blood examinations and some HCV infections. They measure liver function and the degree of infection. The target variable is Hepatitis C stages. Class imbalance was treated in the preprocessing, rendering the dataset convenient for training the predictive ML models.

TABLE I. VARIOUS FEATURES OF THE EMPLOYED DATASET

Feature	Description
WBC	White Blood Cell count - indicates immune system activity.
RBC	Red Blood Cell count - carries oxygen throughout the body.
Plat	Platelets - help with blood clotting
AST 1	Aspartate Aminotransferase - a liver enzyme indicating liver damage
ALT 1	Alanine Aminotransferase - another key liver enzyme
ALT4	ALT after 4 weeks - used to monitor liver treatment response
ALT 12	ALT after 12 weeks - longer-term liver enzyme monitoring
ALT 24	ALT after 24 weeks - extended liver response tracking
ALT 48	ALT after 48 weeks - used for final liver recovery status
RNA Base	Baseline HCV RNA level - initial viral load
RNA 4	HCV RNA level after 4 weeks - early viral response
RNA 12	HCV RNA level after 12 weeks - mid-term treatment effectiveness
RNA EOT	End of Treatment RNA level - confirms treatment success.
RNA EF	RNA level after follow-up - evaluates sustained virologic response
ALT after 24 weeks	ALT enzyme level after 24 weeks - liver recovery marker
ALT 36	ALT after 36 weeks - mid-to-late stage liver enzyme
Age	Age of the patient - a risk factor for disease progression
Baseline histological Grading	Initial liver tissue damage score

Table I shows the clinical and laboratory characteristics used in the project (WBC, RBC, ALT/AST liver enzymes, and diverse RNA levels measured during the treatment time course). These characteristics served as input for a machine learning-based model to predict and classify the histological baseline stage of patients with HCV. Every feature adds some medical context that helps to predict the disease progression accurately.

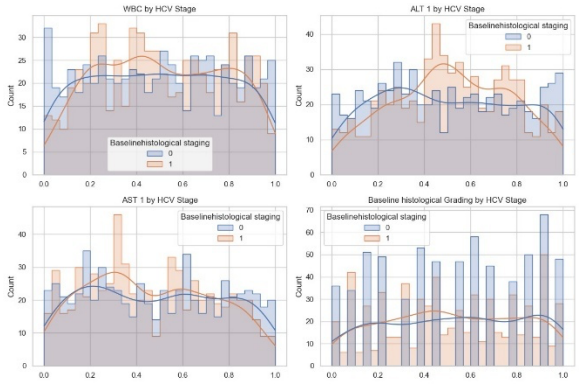

Fig. 1. Histogram plot for WBC, ALT 1, AST 1, and Baseline histological grading features

Fig. 1 represents the distribution of four relevant clinical parameters (WBC, ALT1, AST1, and Baseline histological Grading) in the two stages of HCV (class 0 and class 1) over the Balanced data of the Training set. We compare the densities of these features for each subplot between the two classes. Whereas WBC, ALT1, and AST1 present overlapped feature

distributions with different mass weights, the histological Baseline Grading separation presents an apparent separation, with class 0 showing a higher and spread distribution. These visual distinctions indicate that some features might have higher visual predictability, and we incorporate them for training the classifiers.

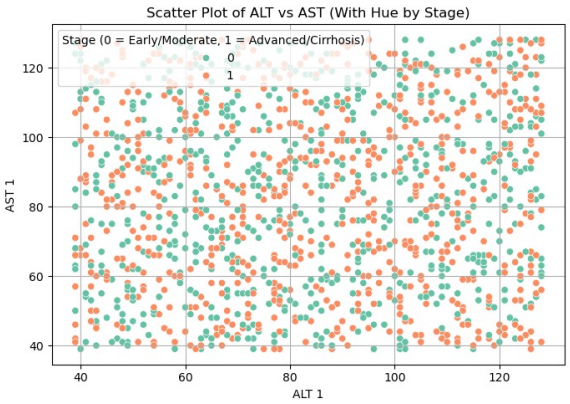

Fig. 2. Scatter plot for AST 1 vs ALT1

Fig. 2 represents the relationship of two enzymes: ALT 1 and AST 1 in HCV-infected patients, divided into two stages of disease (0 = early/Moderate and 1 = Advanced/Cirrhosis). Each point is an individual patient, color-coded by stage. This distribution indicates that similar ALT/AST values are found for patients in class 1 (both early and advanced HCV stages). However, some groups suggest that higher combinations of ALT and AST can more accurately indicate advanced fibrosis. This visualization provides evidence of the relevance of these enzymes as biomarkers for staging HCV when employed in conjunction with machine learning classifiers.

B. Dataset Preprocessing

A structured preprocessing pipeline was applied to the HCV dataset to create the best quality training data for the machine learning model, illustrated in Fig. 3. The first step was to clean the data: dealing with missing values and reusing column titles. When training the model, Baseline histological staging was considered the target variable reflecting the stages of liver fibrosis, and all other clinical and laboratory measurements were subjected to input features.

Normalization was performed through Min-Max Scaling to make the values of all features on the same scale in [0,1]. This was especially crucial for the performance of distance-based models (KNN, SVM) and the convergence of tree-based models (RF/XG Boost).

To reduce dimensionality and enhance generalization, 18 informative features were searched using Recursive Feature Elimination (RFE) with Random Forest as the base estimator. These are essential prognostic parameters, namely, ALT4, ALT12, RNA 4, WBC, PLATELETS, and Baseline histological grading. Feature selection facilitated the exclusion of noisy and irrelevant features, which helped lower the danger of overfitting.

Due to the class imbalance in the target variable, Borderline SMOTE was used to oversample the minority classes. The technique balanced the dataset and suppressed the bias of the majority class by creating new points using the nearest neighbors of minority samples.

The overall preprocessing workflow is illustrated in the flowchart below:

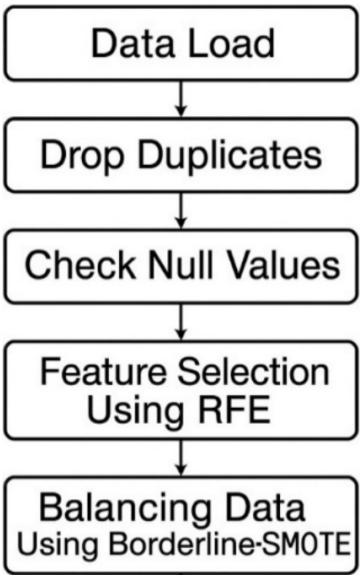

Fig. 3. Dataset preprocessing sequence

Key preprocessing equations, such as the Min-Max Scaling formula and RFE scoring function, are included below for reference:

Min-Max Scaling:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

These preprocessing steps led to a clean, scaled, feature-engineered, and balanced dataset, which played an essential role in the high performance and reliability of predictive models built in this study.

C. Machine Learning Models

This study uses machine learning methods to predict HCV infection from clinical data in a stacked ensemble learning framework. The first layer of the ensemble adopted four base classifiers: Random Forest, XG Boost, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). These models were

trained separately on preprocessed and balanced data with optimized or default hyperparameters.

The **Random Forest (RF)** was trained on a class-balanced dataset with Baseline SMOTE to resolve class imbalance. With 100 estimators and a set random state, the model utilized the strength of multiple decision trees for strong classification (n_estimators=100, random_state=42). It had 67% accuracy on the test set, indicating good generalization capability. However, we could get more improvement by tuning max_depth, min_samples_split, or class_weight, especially on the minority class.

The **XG Boost (Extreme Gradient Boosting)** model was fitted to a class-balanced dataset obtained with BaselineSMOTE to correct for the imbalance between the different stages of HCV infection. For the default model configuration (use_label_encoder=False, eval_metric='logloss', random_state=42), the score is 63% of the real test set. Although it provided mediocre performance, recall for the minority class might improve by tuning hyperparameters (max_depth, n_estimators, learning_rate) or early stopping.

All models were imbalanced using Baseline SMOTE in a **Decision Tree (DT)** model with default hyperparameters (criterion='gini', max_depth=None, min_samples_split=2, random_state=42). It achieved an accuracy of 57.2% on the testing dataset. It performed modestly for class 0 (precision: 0.73, recall: 0.68) with lower sensitivity for class 1 (precision: 0.23, recall: 0.28), indicating the difficulty of predicting minority classes in the medical datasets even after balancing.

A **Support Vector Machine (SVM)** classifier was built using a Baseline SMOTE balanced dataset to overcome the class imbalance. We used the radial basis function (RBF) kernel with standard hyperparameters (kernel='rbf', C=1.0, gamma='scale', random_state=42). The test set accuracy for the model was 57.6%, which performed better in class 0 compared to class 1. Though MSM was able to catch the trends of the general population, it underperformed minority class prediction, indicating hyperparameter optimization/wider ensembles should be considered in imbalanced medical classification tasks.

The **K-Nearest Neighbors (KNN)** model was trained with k = 5 on a Baseline SMOTE balanced dataset. Although a simple and interpretable model, the KNN accuracy on the testing set was only 46%. This poorer performance probably arises from KNN's sensitivity to irrelevant variables and inefficiency under high-dimensional or noisy data. Better results could be obtained with feature selection, normalization, and tuning the n neighbors parameter.

The **Stacked Ensemble** service model has four strong base learners: Random Forest, XG Boost, SVM, KNN, and XG Boost, as a meta-learner. The class probabilities generated in each base model were accumulated and fed to the meta-model for prediction. Borderline SMOTE was used to resample the data, as well as the Min-Max Scaler scaled features. This multi-model technique proved very robust by capturing the strengths of the base models and minimizing the weaknesses of individual classifiers. It achieved a good 88% accuracy on the test set.

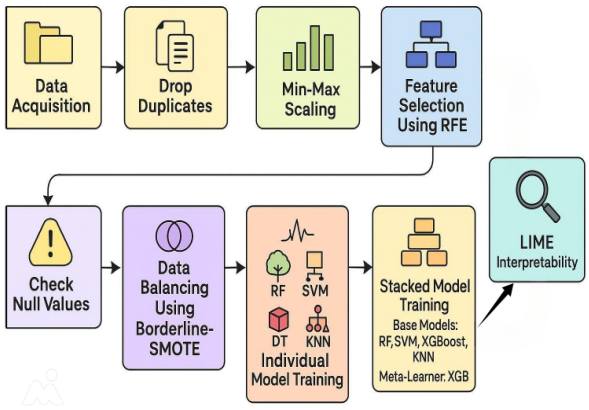

Fig. 4. Working sequences of the proposed Hepatitis C prediction system

Fig. 4 demonstrates the entire machine learning process applied to HCV stage prediction. Starts with step 1: data collection, where the HCV dataset is obtained from a reliable source. The data is then cleaned by removing duplicates and normalizing feature values between 0 and 1 (min-max scaling) to enhance the model performance further. Subsequently, recursive feature elimination (RFE) is performed to select the top significant clinical features for prediction. The stage also deals with null values, so no data is missing and nothing will shift the model's training. In addition, to deal with the class imbalance problem, the Borderline-SMOTE is adopted to generate synthetic samples close to the decision boundary for balancing the dataset and representing each class. After that, single models RF, SVM, DT, and KNN are built on the pre-processed data. These probability predictions are aggregated in a stacked ensemble model (i.e, XG Boost is the meta-learner)to achieve better overall accuracy. LIME (Local Interpretable Model-agnostic Explanations, finally) is used to interpret decisions of the stacked model so that research and clinical staff know which features have the most effect on the predictions.

This subsection describes the performance evaluation metrics for the classification models. These measures are Accuracy, Precision, Recall, and F1-score. These

are crucial in assessing how well the model predicts the positive and negative cases in medical diagnostic tasks, such as Hepatitis C classification.

Accuracy: proportions the correct predictions (true positives and negatives) over the total number of cases evaluated.

$$Accuracy = \frac{\mathrm{PS} + \mathrm{NS}}{\mathrm{PS} + \mathrm{NS} + \mathrm{FS} + \mathrm{IS}} \quad (2)$$

Precision: computes the ratio of the number of accurate optimistic predictions to the number of positive predictions.

$$Precision = \frac{\mathrm{PS}}{\mathrm{PS} + \mathrm{FS}} \quad (3)$$

Recall: It is calculated as the number of true positives divided by the sum of true positives and false negatives.

$$Recall = \frac{PS}{\mathrm{PS} + \mathrm{IS}} \quad (4)$$

F1 Score: is a balance of precision and recall. It's the geometric average of precision and recall.

$$F1_{score} = \frac{2\mathrm{PS}}{2\mathrm{PS} + \mathrm{PS} + \mathrm{IS}} \quad (5)$$

III. Results and Discussion

This study built an ensemble model for HCV prediction using an extensive clinical data set from the UCI machine learning repository. The pre-processing pipeline was feature selection through recursive feature elimination and class balancing, utilising Borderline SMOTE to address the imbalance in the class distribution. The model was an ensemble of four base classifiers, namely Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Xtreme Gradient Boosting (XG Boost), and their predicted class probabilities were used to train a meta-learner with XG Boost.

The stacked ensemble model outperformed single models, with an accuracy of 88% on the test set. Two single classifiers, Random Forest and XG Boost, were found to have the best accuracy, but they only generalized better when the meta-model was used. Performance measures, such as precision, recall, and F1-score, confirmed good performance, especially in detecting minority class subjects that are usually challenging to classify in medical datasets accurately. Moreover, model intelligibility was improved through the LIME framework, producing local explanations for each prediction and improving the proposed system's transparency and credibility for clinical decision support.

TABLE II. HYPERPARAMETER VALUE RANGES FOR ALL THE ML MODELS

Model	Hyperparameter Value Range	Optimized value
SVM	C: [0.1, 1, 10, 100, 1000], gamma: [1, 0.1, 0.01, 0.001, 0.0001], kernel: [linear, RBF]	C:10, gamma:1, Kernel: RBF
Random Forest	n_estimators: [200,500], max_features: [auto,sqrt, log2], max_depth: [4,5,6,7,8], criterion: [Gini, entropy]	n_estimators: 200, max_features: auto, max_depth:10, criterion: entropy
Decision Tree	criterion: [Gini, entropy], max_depth: np.arange(3, 15)	criterion: entropy, max_depth:6
KNN	n_neighbors: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]}, scoring=accuracy, verbose=1	n_neighbors: 1
XGBoost	max_depth: range (2, 10, 1), n_estimators: range (60, 220, 40) learning_rate: [0.1, 0.01, 0.05]	max_depth: 10, n_estimators: 220 learning_rate: 0.01
Proposed Meta Model	max_depth: range (2, 10, 1), n_estimators: range (60, 220, 40) learning_rate: [0.1, 0.01, 0.05]	n_estimators: 220, max_depth:10, learning_rate: 0.5

Table II summarizes the hyperparameter tuning for all machine learning models applied in the project. Below, it shows the range of search space you would choose for some of the key parameters like C, gamma, n_estimators, max_depth, kernel type, etc. of models such as SVM, random forest, decision tree, KNN, etc. After finetuning, the best hyperparameter values are found, with the best value being: for SVM, C=10, gamma=1, kernel='RBF', and the optimized value for the proposed stacking model is: max_depth=30, n_estimator=220, with learning rate=0.5 for the meta-model (XG Boost). This fine-tuning was performed to achieve the best performance of any model during evaluation.

TABLE III. PERFORMANCE METRICS OF VARIOUS ML MODELS WITH DEFAULT HYPERPARAMETERS

Model	Accuracy	Precision	Recall	F1-score
SVM	63.51	0.64	0.64	0.64
Random Forest	67.11	0.60	0.67	0.63
KNN	45.94	0.62	0.46	0.48
Decision Tree	57.20	0.60	0.57	0.58

XG Boost	69.82	0.66	0.70	0.67
Stacking Ensemble Model	**88.45**	**0.88**	**0.88**	**0.88**

Table "Performance Metrics of Different ML Models with Default Hyperparameters" compares different machine learning methods on the Hepatitis C Virus (HCV) classification. Regarding the individual classification models — SVM, Random Forest, K-NN, Decision Tree, and XG Boost — the best performance was obtained by the XG Boost model, which reached an accuracy of 69.82%. Nevertheless, the combined Stacking Ensemble Model, composed of SVM, KNN, Random Forest, and XG Boost as base learners and XG Boost as a meta-learner, outperformed all single models. It performed the best in terms of all the performance metrics—accuracy = 88.45%, precision = 0.88, recall = 0.88, and F1-score = 0.88, indicating its ability to generalize and address intricate characteristics of HCV data. This is a testament to the power of ensemble learning in a task such as biomedical classification that often requires high precision and sensitivity.

TABLE IV. PERFORMANCE METRICS OF VARIOUS ML MODELS WITH OPTIMIZED HYPERPARAMETERS

Model	Accuracy	Precision	Recall	F1-score
SVM	67.57	0.67	0.68	0.68
Random Forest	64.86	0.61	0.65	0.63
KNN	50.00	0.66	0.50	0.53
Decision Tree	52.70	0.61	0.53	0.55
XG Boost	**70.72**	**0.67**	**0.71**	**0.68**

The table "Performance Metrics of Different ML Models with Optimized Hyperparameters" provides an overview of how the predictive performance of individual ML models has been improved by optimizing parameters through fine-tuning applied to each machine learning model of the HCV classification project. On optimization, the XG Boost model maintained its superiority over other models with an accuracy of 70.72%, precision of 0.67, recall of 0.71, and the F1-score of 0.68, again proving that it is the best model for clinical data. Modest improvements over the baseline were achieved by most models, especially SVM, which achieved 67.57% accuracy with an F1-score of 0.68. Even after fine-tuning, the simpler models (KNN, Decision Tree) failed to work well, indicating that complex models or ensemble approaches are better for representing non-linearities and the class imbalance in the HCV dataset.

TABLE V. PERFORMANCE METRICS OF VARIOUS ML MODELS WITH OPTIMIZED HYPERPARAMETERS AND FEATURE SELECTION

Model	Accuracy	Precision	Recall	F1-score
SVM	68.57	0.77	0.68	0.68
Random Forest	65.86	0.65	0.65	0.66
KNN	55.00	0.69	0.59	0.56
Decision Tree	54.70	0.65	0.58	0.59
XG Boost	73.72	0.69	0.75	0.72
Stacking Ensemble Model	**88.45**	**0.88**	**0.88**	**0.88**

The tabulated results in “Performance Metrics of Various ML Models with Optimized Hyperparameters and Feature Selection” demonstrate the added value gained from hyperparameter tuning and feature selection performed via RFE. Regardless of the pooling method, XGB gave the best single model with 73.72% accuracy and an F1-score of 0.72, demonstrating its efficient capture of more complex patterns. But the Stacking Ensemble Model (SVM, KNN, and RF + XG Boost as meta-learner) outperformed all base models and gave excellent results, i.e, 88.45% accuracy, 0.88 precision, 0.88 recall, and 0.88 F1 score. This also reveals that the high performance gain in the classification of HCV stage can be achieved by integrating the models with well-tuned parameters and significant features.

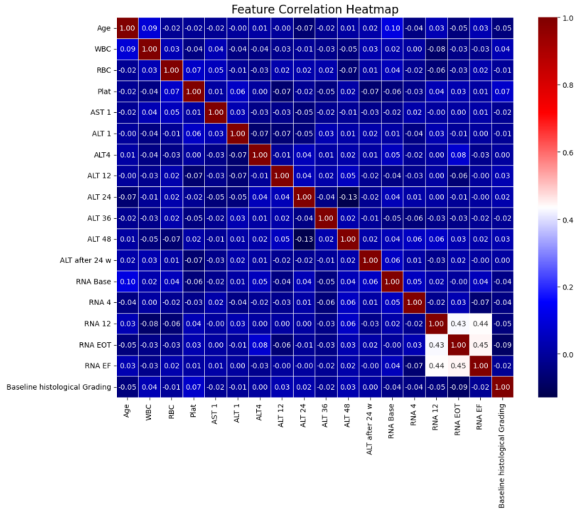

Fig. 5. Heat map diagram of the correlation of various independent variables in the dataset

Fig. 5 provides a visual overview of linearly related features with the target feature ” Baseline histological Grading”. All features carry mostly weak or even close to zero correlation with the target variable, which could imply the lack of a strong independent effect from any one feature. There is also modest intercorrelation among related features; for example, the RNA-based features (RNA Base, RNAT 12, RNA EOT, and RNAEF) are positively correlated, indicating they measure similar aspects of patient condition. Correspondingly, ALT readings at various time points correlate moderately and reflect typical biological progression. Ultimately, the heatmap underlines the intricacy and non-linearity of the data, justifying ensemble models such as the stacker used in this project, which can pick up small patterns involving multiple weakly correlated features.

Fig. 6. Selected features by the feature selection method

Fig. 6 shows relationships and the general distribution of features selected for HCV stage classification in your project. Each matrix diagonal shows the univariate distribution of features for one class (Stage 0 or Stage 1), and the off-diagonal shows a scatter plot of all feature pairs. The scattering of overlapping points shows the non-linear separability, so we need ensemble-unfriendly models like Random Forest and XGBoost. Specific characteristics with somewhat separated distributions would give the classifier information to predict the stage. This display demonstrates that feature and model selection work with your stacked ensemble framework.

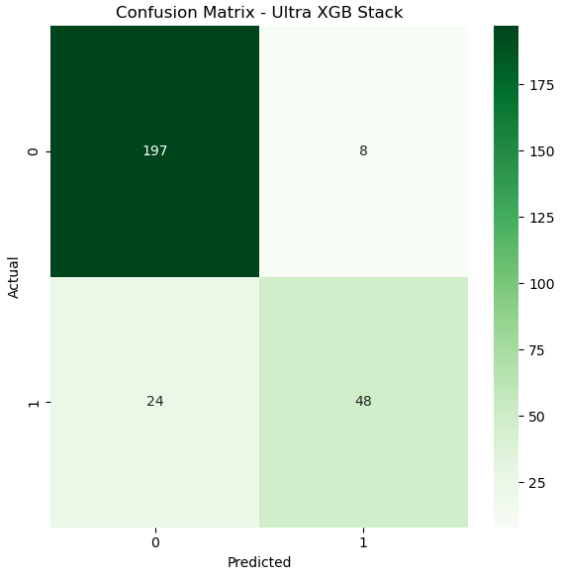

Fig. 7. Confusion matrix of the best-performing stacked ensemble model

The confusion matrix of the Stacked ensemble model demonstrated a model that has shown itself to classify well, with few false positives or false negatives. This indicates that the model has adequate reliability and resistance to distinguish between the classes of Hepatitis C severity.

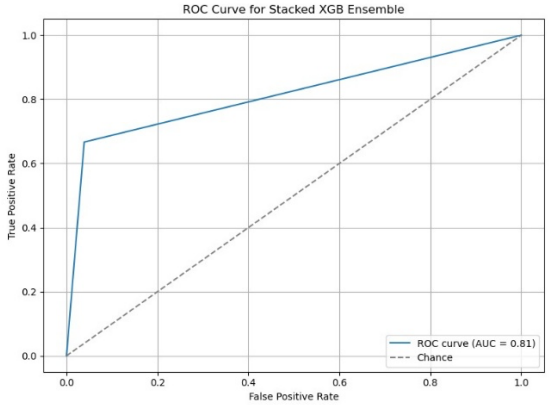

Fig. 8. ROC curve for stacking ensemble model, with AUC = 0.81

ROC curve for the combined XG Boost ensemble, with True Positive Rate (sensitivity) against False Positive Rate. The curve steeply increases to TPR ≈ 0.67 at a relatively low FPR ≈ 0.04, and then increases to (1,1). The Area Under the Curve is 0.81, indicating good general discriminative power of the model, much greater than what would be expected if flipping a coin (the diagonal line).

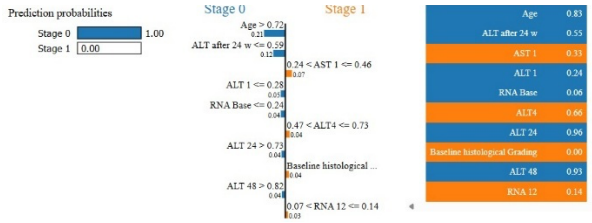

Fig. 9. Machine learning model prediction interpretation by LIME explainable AI library

LIME was applied for interpreting individual predictions. It showed the ones driving more to the prediction, adding interpretability and trust to the model.

TABLE II. COMPARISON OF THE PROPOSED SYSTEM WITH EXISTING WORKS

Ref.	**Model**	**Accuracy/RMSE**	**Other metrics** (Precision)
[1]	HPM	96.82	98.93
[2]	LASSO	97.80	66.70
This work	Stacked Model	88.45	88.00

Table VI compares the proposed Stacked Ensemble Model with LASSO Regression (L1 regularization), and HPM (Hybrid Predictive Model consisting of Improved Random Forest and SVM).

IV. CONCLUSIONS

This paper introduces a machine learning framework that is both interpretable and reliable for early-stage HCV detection. The clinical data set from the UCI Machine Learning Repository we used was preprocessed by removing instances with duplicated cases and missing values and performing RFE for feature selection. Borderline-SMOTE is introduced to address the problem of class imbalance and help the models capture the minority samples better.

A meta-learning model was used using a stacked ensemble learning method, including four classifiers: Random Forest, XG Boost, SVM, KNN, and one meta-learner (XG Boost). The ensemble model achieved good classification accuracy up to 88% and outperformed the traditional single model. Additionally, this LIME algorithm helped enhance the model's interpretability by clarifying the local predictions, thereby making the model more transparent.

The hierarchical ensemble model we proposed exhibited favorable predictability and interpretability, indicating its applicability in clinical decision-making aids. Next, the deep learning models might be incorporated in the analysis, allowing for more sophisticated data dependencies and better predictive performance. In addition, the dataset could be augmented with Electronic Health Record (EHR) data from other sources, and the model could be validated in a clinical and real environment to check its generalizability in healthcare scenarios.

REFERENCES

[1] W. R. Kim, "Global epidemiology and burden of hepatitis C," Microbes and Infection, vol. 4, 2002

[2] G. M. Lauer and B. D. Walker, "Hepatitis C Virus Infection," The New England Journal of Medicine, vol. 345, Jul. 2001

[3] U. K. Lilhore *et al*., "Hybrid model for precise hepatitis-C classification using improved random forest and SVM method," Scientific Reports, vol. 13, 2023.

[4] S.-H. Kim, S.-H. Park, and H. Lee, "Machine learning for predicting hepatitis B or C virus infection in diabetic patients," Scientific Reports, vol. 13, 2023.

[5] UCI Machine Learning Repository, "Hepatitis C Virus (HCV) for Egyptian patients Dataset," [Online]. Available: https://archive.ics.uci.edu/dataset/503/hepatitis+c+virus+hcv+for+egyptian+patients.

[6] S. C. Nandipati, C. XinYing, and K. K. Wah, "Hepatitis C virus (HCV) prediction by machine learning techniques," Applied Modeling and Simulation, vol. 4, 2020.

[7] 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, 20