International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 08 Aug. 2025, pp: 1053-1064 www.ijaem.net ISSN: 2395-5252

A Data-Driven Comparative Study of Machine Learning Algorithms for Hepatitis C Diagnosis

Njoku Perfect Izuchukwu[1], Sule Haruna Sani[1*], Alagbe Sefiu Adebayo[2], Otubobola Seyi Ayomide[3], Samuel Edwin Inalegwu[4], Oduma Elijah Oduma[5], Ato Joy lember[6], Sanyaolu Esther Precision[7], Moses Micheal Oluwapelumi[8], Okon Nkoo Etim[9], Raisul Anwar[10], Bereket Nate Boye[11],Racheal Taiwo Bakare[12], Jonathan Boomni Oye[13], Eneh Joy Chinemerem[14]

1, 5, 6, 9, 10, 11, 12, 13& 14Department of Public health and Evidence-base Medicine, I.M. Sechenov First State Medical University, Moscow.
1, & 2, 3 Department of Statistics, Federal University of Agriculture Abeokuta.*
4 Department of Laboratory Services, Africa Medical Center of Excellence
7 Department of Nutrition and Dietetics, Ladoke Akintola University of Technology, Ogbomoso, Oyo State.
8 Department of Computer Science, Ayede Federal Polytechnic, Oyo State.

Date of Submission: 25-08-2025 Date of Acceptance: 05-09-2025

ABSTRACT
Hepatitis C is a viral infectious disease caused that affects 58 million people globally making it a public health crisis, and it is reported that 20-30% of cases develop to severe liver disease. This creates the need for data-driven based method that such as Machine learning that depends on data inputs in making accurate outputs with high precision. This study explores the use of Machine learning algorithms in optimizing the detection of hepatitis C status of patients. This research utilizes a publicly available dataset that contains 12 explanatory features and one target feature that addresses the patient's diagnosis status. Data preprocessing techniques such as checking of missing values, assessment for possible outliers and Exploratory data analysis was carried on the dataset to improve the quality of the dataset for predictive analysis. The EDA results reviews that majority of the features had significant relationship (p-value < 0.05) with the target variable. Also, we observed the distribution of the disease diagnosis by gender reviews that females showed higher prevalence, and when we consider age, older people are more likely to contract the HCV. This result aligns with research findings from recent studies. The dataset was scaled and separated using a ratio split of 80:20 (80% of data for training). We applied SMOTE algorithm to the training data to balance the diagnosis distribution of the target variable. The model was trained using a wide parameter space using a cross-validation fold of 10 to select the best parameters. The ML models evaluated in this research include Logistic Regression, Support Vector Machine, Random Forest, XGBOOST, Decision Tree, and Extra Trees. The model performance metrics shows that XGBOOST was superior with an accuracy score of 98%, followed by Random Forest (97%), Decision Tree (94%), Extra Trees (93%), SVM (92%), and Logistic Regression (87%). The Kappa statistics for XGBOOST which was estimated as 0.8827 shows that difference between the predicted instances and the actual instances was minimal. The performance metrics of the XGBOOST was also contrast with ML algorithms trained by recent researchers in detecting HCV and it shows our model was more precise underscoring the significance of data-preprocessing, feature-scaling, SMOTE and hyperparameter tuning in predictive analysis in healthcare.

KEYWORD: Machine learning, Hepatitis C, Imbalanced dataset

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 08 Aug. 2025, pp: 1053-1064 www.ijaem.net ISSN: 2395-5252

I. INTRODUCTION

Sugerman (2014) described Hepatitis C as a blood-borne viral infection that affects the liver primarily and it was reported by Safioleas & Manti, (2007) to be associated with diseases such as cirrhosis, and hepatocellular carcinoma. According to Sugerman (2014) study, 20-30% of the reported cases of HCV develop into severe liver disease over time. A reported population of 58 million people is infected with Hepatitis C globally (Mahdyeh et al., 2024). This high global prevalence underscores the need to develop a vaccine for hepatitis C, and based on our findings, there's no vaccine for controlling this disease (Halliday et al., 2011). This creates a gap for early identification of Hepatitis Virus to mitigate its spread and improve patient outcomes. Recent studies highlight the proficiency of Artificial intelligence algorithms in detecting hepatitis C efficiently. Machine learning algorithms have reported to be efficient in predicting obesity risk, diabetes, and kidney related diseases with high precision and accuracy (Sule. H et al., 2025; KM, 2020; Afia et al. 2023). These studies underscore the significance of integrating ML algorithms in optimizing the detection of infectious disease. The primary aim of this study is to explore the efficiency of various ML algorithms which includes Random Forest, XGBOOST, Decision Tree, Extra Trees, Support Vector Machine, and Logistic Regression in predicting the HCV status of patients with high precision and accuracy.

This research structure includes Section 2.0 which focuses on related studies that adopted ML algorithms in optimizing the detection of HCV; Section 3.0 focuses on the description of the proposed research methodology and the metrics used in evaluating the ML algorithms; Section 4.0 discusses the research findings of the EDA and ML algorithms with relation to predicting HCV; Section 5.0 and 6.0 focus on assessing the results with results from recent studies and the research conclusion.

II. RELATED STUDIES

Heba et al. (2023) investigated the performance of four ML algorithms which includes as Naïve Bayes, Logistic Regression, Random Forest (RF), and K-nearest neighbor (KNN) predicting HCV. The study results reviews that Random Forest yielded the highest accuracy of 94.06%, Logistic Regression & Naïve Bayes 93.01%, and KNN 92.66%.Hashem et al. (2020) developed a ML-based framework for predicting hepatocellular carcinoma status of patients. The study utilized Logistic Regression, Decision Tree, and Classification and Regression Tree (CART). The study reported Decision Tree has the best model with an accuracy of 99%. KayvanJoo et al. (2014) applied Decision Tree, Support Vector Machine, Naïve Bayes, and Neural Network on viral nucleotides results in predicting HCV. The authors utilized various feature selection techniques which includes Chi-square, Gini-index, Deviation, Info-Gain, Info-Gain Ratio, SVM, PCA, Uncertainty, Relief, and Rule. The study results reported an average accuracy of 85%. Satish et al. (2020) investigated the efficiency of KNN, NN, SVM, GNB and RF in handling both binary and multiclass prediction of HCV. The study results conclude that KNN attained the highest accuracy of 54.56% in both the multi and binary classification cases.Nahla et al. (2019) applied ML in optimizing the prediction of liver fibrosis. The study observed that Random Forest attained the best prediction results with an AUC score 0.903 (fibrosis), 0.894 (advanced fibrosis) and 0.822 (mild/advanced fibrosis). Mahdyeh et al. (2024) applied SMOTE and feature selection in improving the performance ML models in predicting HCV. The study utilizes ML models like Random Forest, Decision Tree, XGBoost, AdaBoost, and logistic regression. The study results indicated that AdaBoost attained the best performance combining an ANOVA feature selection strategy. Azadeh et al. (2023) applied data mining technique in combination with ML algorithms which includes Support Vector Machine (SVM), K-nearest Neighbors (KNN), Logistic Regression, decision tree, extreme gradient boosting (XGBoost), and Artificial Neural Networks (ANN). The researchers highlighted that SVM and XGBOOST attained the best accuracy. A comparative analysis which integrated ML algorithms Logistic Regression, Random Forest, Decision Tree, C4.5 and Multilayer perceptron classifiers in predicting HCV (N Komal Kumar & D Vigneswari; 2019). The study results shows that Random Forest had the highest predictive accuracy of 90.32%. Umesh et al. (2023) integrated a hybrid RF and SVM in predicting HCV using SVM, DT, RF, BGLM, and MARS. The study assessed the performance ML algorithms using two cross-validation split 5 and 10 to train these models. The proposed method attained its best performance using k=10 of 96.29%.

III. METHODOLOGY

3.1 Data Description

The dataset used in this study is a publicly available dataset on the Kaggle database. The

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 08 Aug. 2025, pp: 1053-1064 www.ijaem.net ISSN: 2395-5252

dataset contains 615 patient records made up of both demographics of the patients, which are age & sex, and the physiological factors Albumin, Alkaline Phosphatase, Alanine Aminotransferase, Aspartate Aminotransferase, Bilirubin, Cholinesterase, Cholesterol, Creatinine, Gamma-Glutamyl Transferase, and Protein. The target variable of the dataset is the hepatitis C diagnosis status, which is distributed into the following categories: Blood Donor, Suspect Blood Donor, Hepatitis, Fibrosis, and Cirrhosis. We recalibrated this hepatitis C diagnosis status into negative and hepatitis C (positive) diagnosis by encoding Blood-Donor and Suspected Blood Donor as negative outcomes, and others were tagged as positive cases.

Table 1.0: Dataset description

Features	**Type**	**Missing Values**
Category (Diagnosis)	Categorical	No
Age	Integer	No
Sex	Binary	No
ALB(Albumin)	Continuous	Yes
ALP(Alkaline Phosphatase)	Continuous	Yes
ALT(Alanine Aminotransferase)	Continuous	Yes
AST(Aspartate Aminotransferase)	Continuous	No
BIL(Bilirubin)	Continuous	No
CHE(Cholinesterase)	Continuous	No
CHOL(Cholesterol)	Continuous	Yes
CREA(Creatinine)	Continuous	No
GGT(Gamma-Glutamyl Transferase)	Continuous	No
PROT (Total Protein)	Continuous	Yes

3.2 Data Preprocessing

Data preprocessing can be described as a strategy used to improve the quality of the dataset so as to extract credible insight from the dataset. It encompasses data handling, missing values, treating outliers, and addressing dataset imbalance (Kommareddy, 2022). In this study, we examined the dataset for missing values and replaced these values using median values. The dataset was reviewed for possible outliers, and the features with outliers were treated by replacing these values with their median values. This was done to improve the quality of the predictive model (Kalisch et al., 2015). The cleaned dataset was analyzed using an exploratory approach, using charts to visualize the trends and relationships between the input features and the target variable. The distribution of diagnosis classes was examined for data imbalance as it poses significant challenges to the model prediction (Altalhan et al., 2025). This issue was addressed using the SMOTE strategy. The predictor variables were scaled using Standardization, and the data set was separated using a split ratio of 80:20 to evaluate the performance level of the ML algorithms selected in this study.

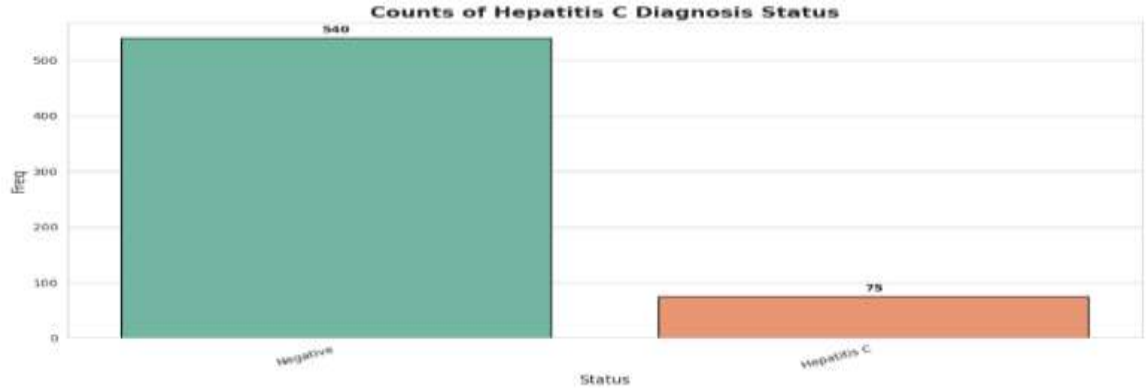

Figure 1.0: Distribution of Hepatitis C diagnosis status

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 08 Aug. 2025, pp: 1053-1064 www.ijaem.net ISSN: 2395-5252

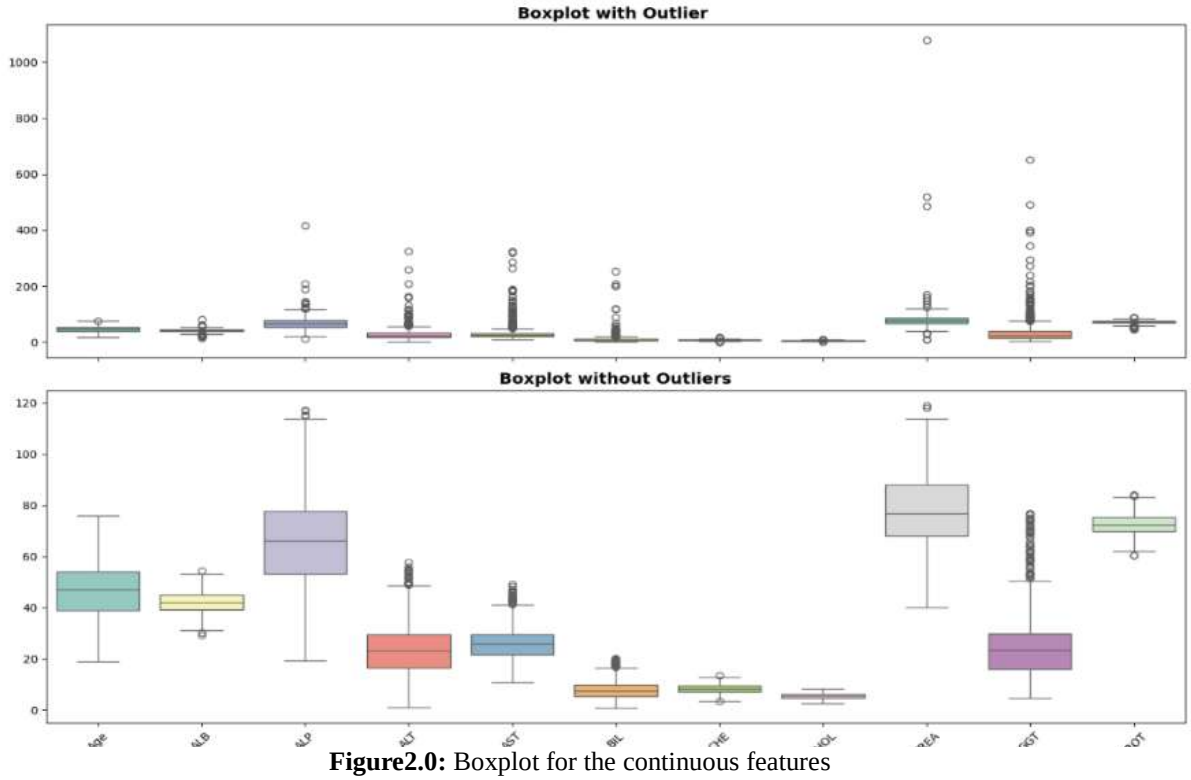

Figure2.0: Boxplot for the continuous features

3.3 Feature Scaling

Feature scaling can be defined as a method used in ML which involves normalization of explanatory variables to improve the outcome of the predictive process (Islam, 2024; Ozsahin et al., 2022). This study utilizes the Standard Scaler Python method in normalizing the input features of the ML algorithms selected. The method utilizes the feature mean and standard deviation in scaling. The mathematical representation of the formula is shown below.

$$x' = \frac{x - \mu}{\sigma} \quad (1.0)$$

Wherex is the observations,μ is the average value of a given feature, and σ is the standard deviation of the feature and x' is the scaled data point.

3.4 Smote Strategy for Class Imbalance

In this study, we utilize SMOTE in addressing the data imbalance of the output feature. This is achieved through a process of linear interpolation based on k-nearest neighbor algorithm (Hairani et al. 2024).This strategy was implemented in this study using the Python imblearn library on the train dataset to prevent issues of data linkage.

3.5 Train-Test Split

Train-Test split is strategy used in ML development process to separate the data of interest into train and test sections for training and evaluating of the generalization of the model (Tan et al. 2021). This study applied an 80:20 train-test split in building the proposed ML algorithms.

3.6 Cross-Validation

Lei Tang (2008) described Cross-validation as statistical method used in development of ML algorithms in separating the data into train and validation sections or folds. Bhagat & Brijesh Bakariya, (2025) stated that the significance of the process helps in ensuring reliability and addresses overfitting. This study applied a cross-validation fold of 10 in developing the proposed ML algorithms.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 08 Aug. 2025, pp: 1053-1064 www.ijaem.net ISSN: 2395-5252

3.7 Machine Learning Algorithms

Table 2.0: Description of the Machine learning algorithms

Algorithm(s)	Model Description
Logistic Regression	Logistic regression can be described as a supervised ML algorithms used specifically for classification related task. This model has been reported to show in good predictive performance in the field of medical research according to Boateng & Abaye (2019) and Makalic & Schmidt (2010).
XGBOOST	This model can be described as a gradient boosting-based model that uses a sparsity aware algorithm, tree learning, and data compression in modelling the output of processes using selected inputs (Fafalios et al., 2020; Chen & Guestrin, 2016).
Random Forest	Random Forest can be defined as a tree-based model that combines the predictive capacity of multiple decision trees using a bootstrapping approach in both regression and classification related task (Liu et al., 2012; Gislason et al., 2004)
Decision Tree	Decision Tree is a tree-based that uses hierarchical structure of nodes and partitioning of data into subspace to create structures to reach the final outcomes at leaf nodes (Babar, 2024).
Extra Trees	Extra Trees can be described as tree-based model which have speed and memory efficiency and as been reported to suitable for dealing with streaming data and medical data (Mastelini et al., 2022; Paul et al., 2024).
Support Vector Machine	Support Vector Machine can define as supervised ML algorithms that uses decision boundaries and planes by maximizing the margins between classes to make predictions (Bharadwaj et al., 2021).

Table 3.0: Machine learning algorithms Hyperparameters space

Model(s)	Hyperparameters and Their Value Ranges	Optimal parameters
Support Vector Machine	C: [0.0001, 0.001, 0.01, 0.1], kernel: ['linear', 'rbf'], gamma: ['scale', 'auto']	C: 0.1, gamma: 'scale', kernel: 'rbf'
Random Forest	n_estimators: [100, 150, 200, 250], max depth: [None, 5, 6, 8, 10], min samples split: [2, 5, 8], min samples leaf: [1, 2], max_features: ['sqrt', 'log2']	max depth: 10, max features: 'sqrt', min samples_leaf: 2, min_samples split': 2, n_estimators: 250
XGBOOST	n_estimators: [50, 100, 150, 200], max depth: [3, 5, 8, 10], learning rate: [0.0001, 0.001, 0.01, 0.1], subsample: [0.8, 1.0], colsample bytree': [0.8, 1.0]	colsample bytree: 0.8, learning rate: 0.1, max depth: 3, n_estimators: 100, subsample': 0.8
Decision Tree	criterion: ['gini', 'entropy'], max depth: [None, 5, 10, 20], min samples split: [2, 5, 10], min samples leaf: [1, 2, 4]	criterion: 'gini', maxdepth: None, min samples leaf': 1, min samples split: 5
Logistic Regression	penalty: ['l1', 'l2'], C: [0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 50]	C: 0.1, penalty: 'l2'
Extra Tree	n_estimators: [100, 150, 200, 250], criterion: ['gini', 'entropy'], max depth: [None, 10, 20], min samples split: [2, 5], min samples leaf: [1, 2]	criterion: 'entropy', max depth: 20, min samples leaf: 2, min samples split: 2, n_estimators: 200

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 08 Aug. 2025, pp: 1053-1064 www.ijaem.net ISSN: 2395-5252

3.8 Model Evaluation

Table4.0: Model performance metrics

Metrics	Description	Formula
Accuracy	Accuracy can be defined as the fraction of instances (both positive and negative) captured correctly by model.	$\frac{TP + TN}{TP + FP + TN + FN}$
Precision	Precision can be defined as the ratio of positive instances retrieved by the model.	$\frac{TP}{TP + FP}$
Recall	Recall can be described as the ratio of positive instances captured accurately by the system.	$\frac{TP}{TP + FN}$
F1-score	F1-score can be described as the harmonic mean of the precision and recall.	$2 * \frac{Precision * Recall}{(Precision + Recall)}$
Kappa	Kappa statistic is used to measure the chance agreement between the predicted outcome and the actual outcome.	

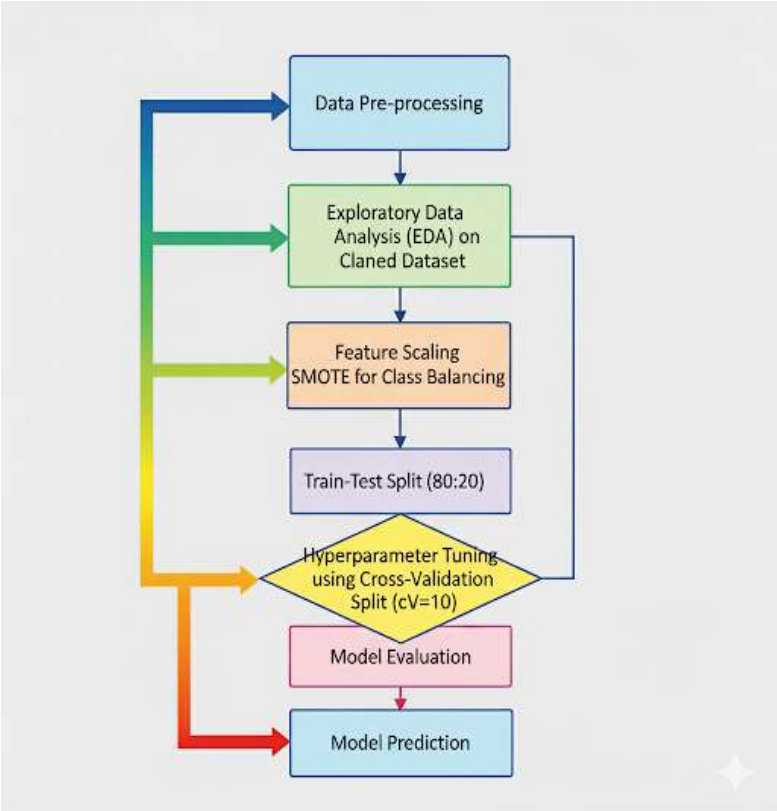

Figure 3.0: Machine Learning Model Development Workflow

IV. RESULT FINDINGS

This section focuses on the demographic distribution of hepatitis C status of patients, relationship between the input features and the target variable and the predictive performance metrics of the ML algorithms evaluated in this study.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 08 Aug. 2025, pp: 1053-1064 www.ijaem.net ISSN: 2395-5252

Figure 4.0: Distribution of Hepatitis C Status by Demographic -Characteristics and Correlation analysis

Figure 4.0 shows the Distribution of Hepatitis C Status by patients' demographic characteristics and the correlation analysis between the predictor variables and the target variable. The plot shows that the distribution of hepatitis C by gender indicated a higher prevalence in females compared to males (Female (53) vs male (22)), indicating that female patients surveyed in this dataset showed higher prevalence. The distribution of hepatitis C by age suggested that hepatitis C has a higher prevalence in older patients compared to younger patients. However, the correlation analysis reveals that only gender had a significant relationship (p-value < 0.05) with the target variable. Also, the physiological factors utilized in this study which includes Albumin (0.46), Alanine Amino transferase (0.22), Bilirubin (0.16), Cholinesterase (0.23), Cholesterol (0.09), Creatinin (0.12), Gamma-Glutamyl Transferase (0.23), and Total Protein (0.09) all had a significant (p-value < 0.05) positive relationship with the target variable. Albumin had the strongest relationship with the target, suggesting it has the highest contributory role in determining the patient's hepatitis C diagnosis status. It was also observed that ALP (0.03, p-value > 0.05) was the only predictor variable, alongside age, that did not show a significant relationship with the target variable.

Table5.0: Machine Learning algorithms performance metrics

Model(s)	**Accuracy**	**Recall**	**Precision**	**F1-score**	**Kappa Statistic**
SVM	0.92	0.87	0.80	0.83	0.6595
XGBOOST	**0.98**	**0.93**	**0.96**	**0.94**	**0.8827**
Random Forest	0.97	0.90	0.95	0.92	0.8389
Decision Tree	0.94	0.88	0.86	0.87	0.7417
Logistic Regression	0.87	0.84	0.72	0.76	0.5281
Extra Trees	0.93	0.81	0.83	0.82	0.6482

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 08 Aug. 2025, pp: 1053-1064 www.ijaem.net ISSN: 2395-5252

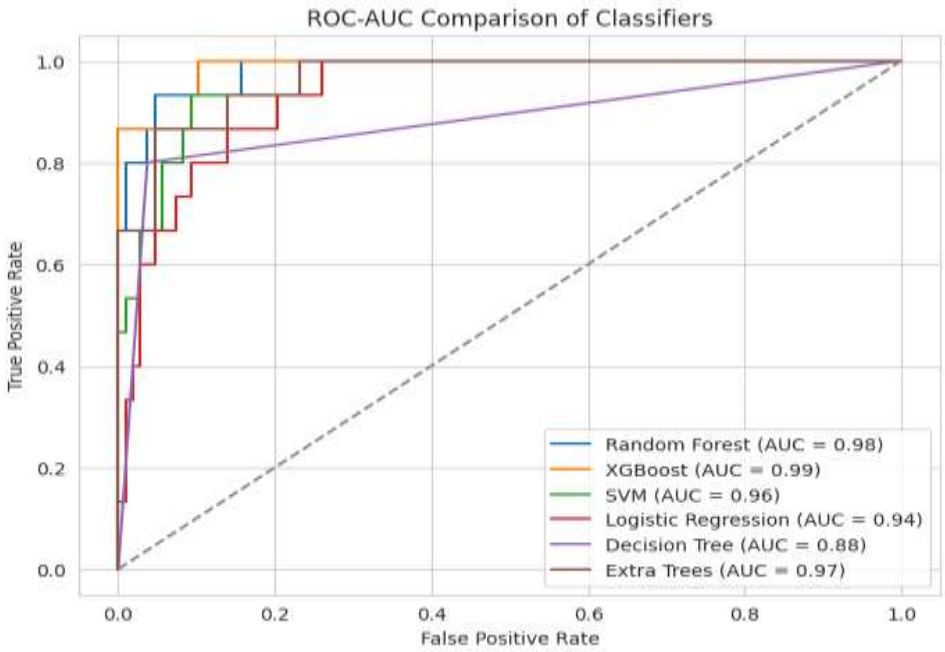

Figure 5.0: ROC-AUC plot

Table 5.0 shows the performance metrics of the Machine learning algorithms considered in this study. The results show that XGBOOST was the optimal model with an accuracy of 0.98, a recall of 0.93, a precision of 0.96, an F1-score of 0.94, and a kappa statistic of 0.8827. The high kappa statistic of the model shows minimal deviation between the predicted diagnosis outcome and the actual outcome. The Random Forest model also showed a significant accuracy of 0.97, a recall of 0.90, a precision of 0.95, an f1-score of 0.92, and a kappa statistic of 0.8389. This result reflects the strong predictive capacity of XGBOOST and Random Forest in determining the Hepatitis Disease status of patients. Decision Tree and Support Vector Machine (SVM) both attained an accuracy of (0.94, 0.92), recall of (0.88, 0.87), and precision of (0.86, 0.80). However, the Decision Tree outperforms the Support Vector Machine in terms of Kappa statistics (0.7417 vs 0.6595). The Extra Trees model outperformed the Support Vector Machine and Logistic Regression models in terms of a higher accuracy (0.93) and Kappa statistic (0.6482). Figure 5.0 shows the ROC-AUC plots, and it indicates the superiority of XGBOOST over the other algorithms based on the model AUC value (0.99). Figure 3.0 shows the confusion matrix over the test dataset of the Machine learning algorithms, and it shows that XGBOOST and the Random Forest model dominate in capturing the actual instances with minimal error.

Table6.0: Best Model (XGBOOST) performance metrics

	Recall	**Precision**	**F1-score**
0	0.98	0.99	0.99
1	0.93	0.87	0.90

Table 6.0 shows the performance metrics of the best-performing model (XGBOOST). The results shows that the negative cases had a strong recall = 0.98, precision = 0.99, f1-score = 0.99, while the prediction of the hepatitis C cases had a recall score of 0.93, precision of 0.87 and f1-score of 0.90. These results suggested a reduced precision in detecting the positive cases while minimizing false negative which crucial in health predictive analysis (Ahamd et al. 2019; Liu et al. 2019). In summary, these results show the superiority of ensemble-based model in predicting Hepatitis C status with high accuracy and reliability.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 08 Aug. 2025, pp: 1053-1064 www.ijaem.net ISSN: 2395-5252

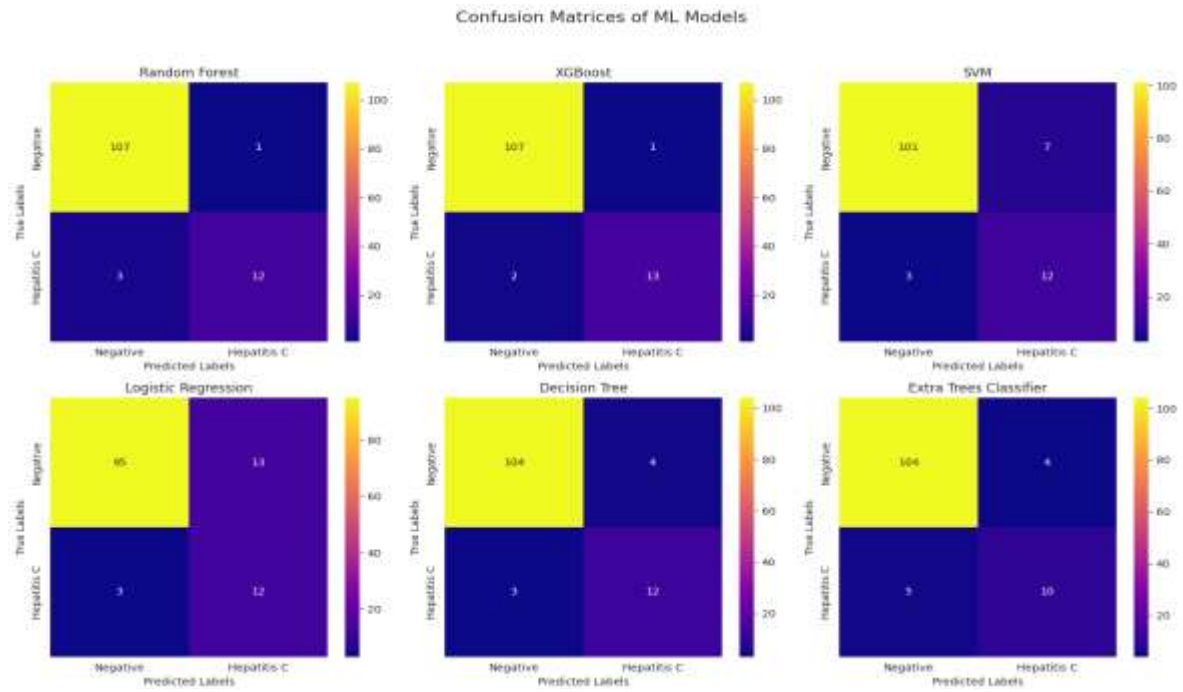

Figure 6.0: Confusion Matrix of the Machine learning algorithms

Table 7.0: Comparative performance with Recent Studies

Author(s)	**Algorithm(s)**	**Best-Accuracy**
Ali MohdAli et al. (2023)	(Artificial Neural Network, K-Nearest Neighbor, Random Forest, Decision Tree, Logistic Regression) + SMOTE	KNN (83.1%)
Anjuman et al. (2024)	Logistic Regression, Support Vector Machine (RBF Kernel), Decision Tree	SVM (RBF Kernel) 92%
Dritsas and Trigka (2023)	Voting classifier with 10-fold cross-validation and SMOTE	80.1%
Umesh et al. (2023)	SVM, DT, RF, BGLM, MARS+ SMOTE	96.29%
Proposed Method	Data preprocessing + feature-scaling + SMOTE+ hyperparameter tuning+ ML algorithms	XGBOOST (98%)

V. DISCUSSION

This study's main focus was to investigate the significance of integrating ML algorithms in the optimization of HCV diagnosis. The study's statistical analysis reveals a higher prevalence of HCV in females compared to males. This outcome aligns with findings that attribute this to female who inject drugs into their system being more likely to contract HCV. HCV distribution by age shows higher prevalence in older patients, indicating that older patients are prone to liver-related diseases that younger patients (Van Dooren et al., 2014; Alvarez et al., 2016). The causes of diseases in older patients could be attributed to factors such as hepatic and extrahepatic comorbidities, including diabetes, cardiovascular diseases, and neurocognitive impairment (Reid et al., 2017). The correlational analysis reveals that all the predictor variables had a significant relationship (p-value < 0.05) with the hepatitis C disease outcome, except age and ALP. Albumin was observed as the variable that showed the contributory effect in determining the outcome of the hepatitis C status of patients, making it a significant predictor of liver-related complications and death in chronic hepatitis C patients (Bonis et

al., 1999). The Machine learning algorithms were trained using a ratio split of 80:20 using SMOTE to balance the target variable and a cross-validation split fold of 10 to train the model and select the optimal parameters. The table shows the optimal parameter attained for each model using a CV of 10. The performance metrics of the ML algorithms show that XGBOOST and Random Forest yielded the best performance based on the accuracy (98%, 97%) and kappa statistics (0.8827, 0.8827). Decision tree, Extra Trees, and SVM had an accuracy of (94%, 93%, 92%) with a precision score of (86%, 83%, 80%). The model's kappa statistic was estimated as (0.7417, 0.6482, 0.6595). The Logistic Regression model had the lowest accuracy score of 86% with a kappa statistic lower than 0.6. Table 6.0 shows the XGBOOST model summary report, and it indicates that the model prediction is robust and generalized well over the unseen dataset. The model performance in this study was compared with that of recent studies. The results indicate that our proposed model performed well with an accuracy score of 98%. This shows the effectiveness of both tree-based algorithms, particularly XGBOOST, in bridging the gap of optimizing early detection of hepatitis C with high precision and accuracy.

VI. CONCLUSION

This study evaluates the proficiency levels of different ML algorithms in predicting Hepatitis C. The study results review that hepatitis C diagnosis distribution by sex shows that females have higher chances of contracting the disease, while the diagnosis distribution by age shows that older people have higher prevalence levels. The proposed method used in this study underscores the significance of addressing data preprocessing, class imbalance using SMOTE, and hyperparameter tuning selection of the optimal parameters for improving the outcomes of predictive models. The proposed models yielded an accuracy score within the range of 88%-98% with XGBOOST having the highest accuracy score of 98%, recall of 93%, precision of 96%, and f1-score of 94%. Also, the model kappa statistics further ascertain the robustness of XGBOOST over the other models, with a value of 0.8827 showing minimal deviation between the predicted values and the actual values. This model, when compared with those of recent studies by Ali M. et al. (2023), Anjuman et al. (2024)& Umesh et al. (2023), shows our model superiority in terms of accuracy. This further shows XGBOOST's superiority over models like SVM, Decision Trees, Logistic Regression, Extra Trees, Artificial Neural Network, K-Nearest Neighbor, and Random Forest. We recommend that future studies explore the use of Generative Artificial Intelligence models in bridging the gap in developing personalized systems for addressing patients with hepatitis C.

REFERENCES

[1]. E. Boateng, D. Abaye (2019). A Review of the Logistic Regression Model with Emphasis on Medical Research
[2]. Tianqi Chen, Carlos Guestrin (2016). XGBoost: A Scalable Tree Boosting System
[3]. S. Fafalios, Pavlos Charonyktakis, I. Tsamardinos (2020). Gradient Boosting Trees
[4]. Zeravan Arif Ali, Ziyad H. Abduljabbar, Hanan A. Tahir, Amira Bibo Sallow, Saman M. Almufti (2023). eXtreme Gradient Boosting Algorithm with Machine Learning: a Review
[5]. Jimin Tan, Jianan Yang, Sai Wu, Gang Chen, Jake Zhao (2021). A critical look at the current train/test split in machine learning
[6]. Manahel Altalhan, Abdulmohsen Algarni, M. Turki-Hadj Alouane (2025). Imbalanced Data Problem in Machine Learning: A Review
[7]. Sukhdevsinh Dhummad (2025). The Imperative of Exploratory Data Analysis in Machine Learning. Retrieved from: https://saspublishers.com/article/21418/
[8]. Mateusz Kalisch, M. Michalak, M. Sikora, Lukasz Wróbel, P. Przystałka (2015). Influence of Outliers Introduction on Predictive Models Quality
[9]. Naga Jahnavi Kommareddy (2022). Data Pre-Processing Framework for Supervised Machine
[10]. Anjuman Ara, Anhar Sami, Daniel Lucky Michael, Ehsan Bazgir and Pabitra Mandal (2024). Hepatitis C prediction using SVM, logistic regression and decision tree. Retrieve from:https://www.researchgate.net/publication/381001142_Hepatitis_C_prediction_using_SVM_logistic_regression_and_decision_tree
[11]. K. van Dooren, S. Kinner, M. Hellard (2014). A Comparison of Risk Factors for Hepatitis C Among Young and Older Adult Prisoners

[12]. J. Halliday, P. Klenerman, E. Barnes (2011). Vaccination for hepatitis C virus: closing in on an evasive target
[13]. A.Gramenzi, F. Conti, F. Felline, C. Cursaro, A. Riili, M. Salerno, S. Gitto, L. Micco, A. Scuteri, Pietro Andreone, Mauro Bernardi (2010). Hepatitis C Virus-related chronic liver disease in elderly patients: an Italian cross-sectional study
[14]. G. Dierckx (2006). Logistic Regression Model
[15]. Kimberly J. Alvarez, A. Smaldone, E. Larson (2016). Burden of Hepatitis C Virus Infection Among Older Adults in Long-Term Care Settings: A Systematic Review of the Literature and Meta-Analysis.
[16]. Michael Reid1, Jennifer C. Price1, and Phyllis C. Tien (2017). Hepatitis C Virus Infection in the Older Patient. Retrieved from: https://pmc.ncbi.nlm.nih.gov/articles/PMC5675560/
[17]. A.Esmaeili, A. Mirzazadeh, G. M. Carter, A. Esmaeili, B. Hajarizadeh, H. Sacks (2017). Higher incidence of HCV in females compared to males who inject drugs: A systematic review and meta-analysis.
[18]. Heba Mamdouh Farghaly, Mahmoud Y. Shams, Tarek AbdEl-Hafeez (2023). Hepatitis C Virus prediction based on machine learning framework: areal-world case study in Egypt. Retrieved from: https://link.springer.com/content/pdf/10.1007/s10115-023-01851-4.pdf
[19]. Hashem S, ElHefnawi M, Habashy S et al (2020) Machine learning prediction models for diagnosing hepatocellular carcinoma with HCV-related chronic liver disease. Comput Methods Programs Biomed 196:105551
[20]. KayvanJoo AH, Ebrahimi M, Haqshenas G (2014) Prediction of hepatitis C virus interferon/ribavirin therapy outcome based on viral nucleotide attributes using machine learning algorithms. BMC Res Notes 7:1–11
[21]. Satish CR Nandipati, Chew XinYing, Khaw Khai Wah (2020). Hepatitis C Virus (HCV) Prediction by Machine Learning Techniques.
[22]. Nahla H. Barakat, Sana H. Barakat, Nadia Ahmed (2019). Prediction and Staging of Hepatic Fibrosis in Children with Hepatitis C Virus: A Machine Learning Approach. Retrieved from: https://synapse.koreamed.org/upload/synapsedata/pdfdata/1088hir/hir-25-173.pdf
[23]. Mahdyeh Abbasi Hezari, Marzieh Baes, Atyeh Abassi Hezari, Mobina Hassanbabaei (2024). Advanced Predictive Modeling for Hepatitis C Diagnosis Using Machine
[24]. Learning. Retrieved from: https://www.researchgate.net/publication/385961651_Advanced_Predictive_Modeling_for_Hepatitis_C_Diagnosis_Using_Machine_Learning
[25]. Azadeh Alizargar, Yang-Lang Chang and Tan-Hsu Tan (2023). Performance Comparison of Machine Learning Approaches on Hepatitis C Prediction Employing Data Mining Techniques. Retrieved from: https://www.mdpi.com/2306-5354/10/4/481
[26]. N Komal Kumar, D Vigneswari (2019). Hepatitis- Infectious Disease Prediction using Classification Algorithms. Retrieved from: https://www.researchgate.net/publication/334956658_Hepatitis-_Infectious_Disease_Prediction_using_Classification_Algorithms
[27]. Ali MohdAli, MohammadR.Hassan,AhmadAl-Qerem, Faisal Aburub, Mohammad Alauthman, Issam Jebreen and Ahmad Nabot (2023). Explainable Machine Learning Approach for Hepatitis C Diagnosis Using SFS Feature Selection. Retrieved from: https://www.mdpi.com/2075-1702/11/3/391
[28]. Hairani Hairani, Triyanna Widiyaningtyas, Didik Dwi Prasetya (2024). Addressing Class Imbalance of Health Data: A Systematic Literature Review on Modified Synthetic Minority Oversampling Technique (SMOTE) Strategies
[29]. Bhagat, Brijesh Bakariya, (2025). A Comprehensive Review of Cross-Validation Techniques in Machine Learning
[30]. E. Makalic, D. Schmidt (2010). Review of Modern Logistic Regression Methods with Application to Small and Medium Sample Size Problems

[31]. Yanli Liu, Yourong Wang, Jian Zhang (2012). New Machine Learning Algorithm: Random Forest
[32]. P. Gislason, J. Benediktsson, J. Sveinsson (2004). Random Forest classification of multisource remote sensing and geographic data
[33]. V. Svetnik, Andy Liaw, Christopher Tong, J. Culberson, R. Sheridan, B. Feuston (2003). Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling
[34]. N. Horning (2010). Random Forests: An algorithm for image classification and generation of continuous fields data sets
[35]. Kiran Nitin Babar (2024). Performance Evaluation of Decision Trees with Machine Learning Algorithm
[36]. Arundhati Navada, A. N. Ansari, Siddharth Patil, B. Sonkamble (2011). Overview of use of decision tree algorithms in machine learning
[37]. Ibomoiye Domor Mienye, N. Jere (2024). A Survey of Decision Trees: Concepts, Algorithms, and Applications
[38]. Sosmita Paul, Geet Bhatt, Aditya Dey (2024). Insights into Ocular Disease Recognition: A Machine Learning Approach
[39]. Umesh Kumar Lilhore, Poongodi Manoharan, Jasminder Kaur Sandhu, Sarita Simaiya, Surjeet Dalal, Abdullah M. Baqasah, Majed Alsafyani, Roobaea Alroobaea , Ismail Keshta & Kaamran Raahemifar (2023). Hybrid model for precise hepatitis-C classification using improved random forest and SVM method. Retrieved from: https://www.nature.com/articles/s41598-023-36605-3.pdf
[40]. KM Jyoti Rani (2020). Diabetes Prediction Using Machine Learning. Retrieved from:https://www.researchgate.net/publication/347091823_Diabetes_Prediction_Using_Machine_Learning
[41]. Afia Farjana, Fatema Tabassum Liza, Parth Pratim Pandit, Madhab Chandra Das, Mahadi Hasan, Fariha Tabassum, Md. Helal Hossen (2023). Predicting Chronic Kidney Disease Using Machine
[42]. Learning Algorithms. Retrieved from: (PDF) Predicting Chronic Kidney Disease Using Machine Learning Algorithms