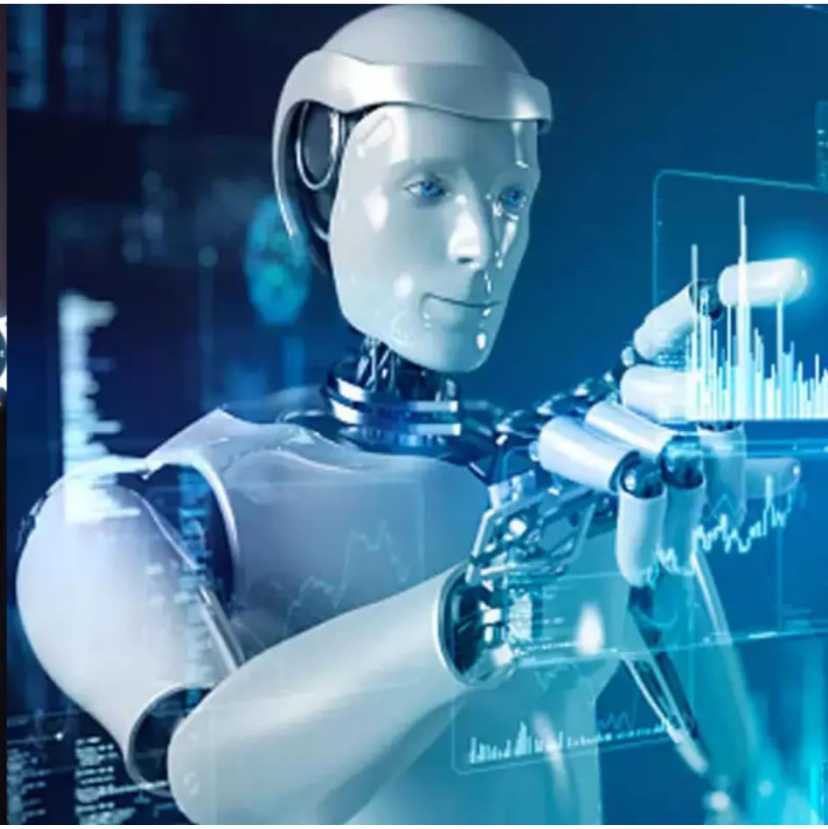
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 4, April 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404382|

Adaptive Data Pre-Processing and Ensemble Modelling for Multiclass Liver Disease Prediction

Yasaswi Ramya Sanapala[1], Dr. Ch. Ramadevi[2]

M.Tech Student, Sir C R Reddy College of Engineering, Eluru, India[1]

Associate Professor, Department of CSE, Sir C R Reddy College of Engineering, Eluru, India[2]

ABSTRACT: Liver disease encompasses a spectrum of health conditions caused by various factors that impair liver function over time. Effective management of Hepatitis C, a prevalent liver disease, relies on accurate and timely prediction of risk factors and disease progression across stages such as fibrosis and cirrhosis. This research proposes a comprehensive framework for multiclass liver disease prediction using the Hepatitis C dataset from the UCI repository. An adaptive data preprocessing technique is introduced to enhance model efficacy, incorporating class-specific mean imputation, outlier rejection, log normalization, feature selection, feature scaling, and data balancing. The preprocessing approach underwent thorough evaluation, demonstrating significant performance improvements. Additionally, ensemble models combining multiple machine learning classifiers were proposed to further boost prediction accuracy. After rigorous parameter optimization, the best-performing model achieved exceptional accuracy, with 99.87% on the training set and 99.80% on the test set, surpassing previous studies. To facilitate practical application, a user-friendly interface was developed, allowing medical professionals to input patient data and promptly receive risk assessments for liver disease, supporting informed clinical decision-making.

KEYWORDS: Machine Learning, Liver Disease, Hepatitis C Virus, Electronic Health Records, Data Preprocessing, Multiclass Classification, Ensemble Model.

I. INTRODUCTION

In recent decades, the significant advancements in machine learning have enabled its widespread application in healthcare, ranging from disease diagnosis to patient prognosis prediction. Machine learning has become an invaluable tool in medical diagnostics, offering the capability to analyze vast amounts of data, uncover patterns, make accurate predictions, and continually enhance its performance over time [1]. Models such as Decision Trees (DT), Support Vector Machines (SVM), Random Forests (RF), Logistic Regression (LR), and Neural Networks have demonstrated promising results in various medical applications [2].

This paper focuses on liver disease, a major global health concern due to its high prevalence, severe complications, and impact on mortality rates. Approximately 2 million people die annually from liver disease, with around 1 million deaths attributed to cirrhosis-related complications and the remaining cases linked to viral hepatitis and hepatocellular carcinoma (HCC) [3], [4]. The liver, a vital organ, is responsible for detoxifying metabolites, producing digestive enzymes, synthesizing proteins, regulating metabolism, managing red blood cells (RBCs), and storing glucose [5], [6].

Liver disease can arise from various factors, including viral infections, alcohol consumption, and obesity. Hepatitis A, B, C, D, and E are common viruses that cause liver infections. Excessive alcohol intake or fat accumulation in the liver may lead to conditions like alcoholic hepatitis, fatty liver disease, and cirrhosis. Additionally, unhealthy lifestyle choices, such as smoking and poor dietary habits, can accelerate liver damage.

In this study, we focus on Hepatitis C, fibrosis, and cirrhosis using the Hepatitis C dataset from the UCI repository. Hepatitis C is a viral infection caused by the Hepatitis C Virus (HCV) and is primarily transmitted through blood-to-blood contact, including unsafe blood transfusions, organ transplants, and, less commonly, sexual contact [7]. While antiviral medications can effectively treat HCV, early diagnosis is crucial to prevent disease progression.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404382|

Fibrosis is an intermediate stage of liver disease characterized by the accumulation of extracellular matrix (ECM) proteins. Often asymptomatic, fibrosis frequently remains undiagnosed until it advances to cirrhosis [8]. Cirrhosis represents the final stage of chronic liver disease, marked by irreversible scarring and loss of liver function. Without timely intervention, cirrhosis may lead to life-threatening complications, including liver failure and hepatocellular carcinoma, the most common form of liver cancer [9].

This paper proposes a comprehensive framework for multiclass liver disease prediction by leveraging machine learning techniques. Through adaptive data preprocessing and ensemble modeling, we aim to achieve accurate classification of liver disease stages, facilitating early diagnosis and effective treatment planning.

II. RELATED WORKS

Numerous studies have explored machine learning approaches for liver disease prediction, employing various models and data preprocessing techniques. This section reviews relevant research to contextualize our proposed methodology and highlight its advantages.

Farghaly et al. conducted a real-world case study using machine learning models to predict Hepatitis C Virus (HCV) infections among healthcare workers in Egypt. By applying classifiers like Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF), and Logistic Regression (LR), the study achieved a maximum accuracy of 94.88%, with RF as the most effective model [10]. Similarly, Ahammed et al. implemented a machine learning model for classifying liver disease stages in HCV patients using 29 distinct features. Through filter-based feature selection, the study found that the K-Nearest Neighbors (KNN) algorithm performed best, followed by RF and SVM [11].

Ayeldeen et al. developed a six-stage model, which included data preprocessing, feature selection, and classification using a Decision Tree (DT) model. With an accuracy of 93.7%, their study highlighted the benefits of selecting the most relevant features to reduce complexity [12]. In another study, Amin et al. introduced an integrated projection-based statistical feature extraction technique using Principal Component Analysis (PCA), Factor Analysis (FA), and Linear Discriminant Analysis (LDA). Although their ensemble classifier did not yield the highest accuracy, it demonstrated effectiveness in scenarios with large datasets requiring dimensionality reduction [13].

Islam et al. applied DT, NB, KNN, RF, and Deep Learning to the same Hepatitis C dataset used in our study. Their results showed Deep Learning as the most accurate model (95.50%), followed by RF with an accuracy of 94.29% [14]. Verma et al. addressed class imbalance using the Minimum Redundancy Maximum Relevance (MRMR) algorithm for feature selection. Their optimizable KNN model achieved a remarkable 100% accuracy on the testing dataset [15].

Mostafa et al. proposed a liver disease prediction model using SVM with Particle Swarm Optimization (PSO) for enhanced accuracy. The study applied Multiple Imputation by Chain Equations (MICE) for missing data management and used PCA for dimensionality reduction [16]. Li et al. developed a hybrid approach using Random Forest, Logistic Regression, and the Artificial Bee Colony (ABC) algorithm for parameter optimization, demonstrating the effectiveness of model combinations [17].

Furthermore, Sivasangari et al. evaluated the performance of SVM, DT, and RF using the Indian Liver Patient Dataset (ILPD), with SVM emerging as the most accurate model at 95.18% [18]. Ali et al. expanded this approach for heart disease prediction using preprocessing methods like Interquartile Range (IQR) for outlier removal. Their study found that KNN, RF, and DT performed exceptionally well, achieving 100% accuracy [19].

Despite the advancements in liver disease prediction, the existing studies exhibit notable limitations:
Inefficient Data Preprocessing: Many studies relied on basic preprocessing techniques without addressing data imbalance, outlier management, or log normalization, leading to suboptimal model performance.
Limited Model Exploration: Most studies focused on individual classifiers without exploring ensemble models that could enhance prediction accuracy.
Lack of Robust Evaluation: Insufficient efforts were made to ensure balanced test datasets, potentially introducing bias during model evaluation.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404382|

To address the identified gaps, this paper proposes a novel approach to liver disease prediction by integrating advanced data preprocessing techniques with machine learning models. The key contributions of our research are as follows:

1. **Adaptive Data Preprocessing**: We introduce an adaptive data preprocessing framework that tailors preprocessing steps based on dataset characteristics. The approach includes class-specific mean imputation, outlier rejection, log normalization, feature selection, feature scaling, and data balancing. This significantly enhances the performance of baseline classifiers without requiring complex models.
2. **Ensemble Model Development**: In addition to improving individual classifier performance, we propose a set of ensemble models by combining base classifiers. This ensures robust and reliable predictions across different liver disease stages.
3. **User-Friendly Interface**: To facilitate practical application, we design an intuitive user interface for medical professionals. The system allows users to input patient data easily and receive real-time liver disease predictions, contributing to efficient clinical decision-making.

III. METHODOLOGY

The block diagram of the system architecture is depicted in Figure 1, highlighting the key stages of our model. The following is a concise description of each stage:

1. **Data Extraction**: The dataset was sourced from the UCI HCV repository.
2. **Data Preprocessing**: Our adaptive data preparation technique was applied to clean and preprocess the dataset, ensuring its suitability for machine learning model development.
3. **Data Splitting**: The preprocessed data was partitioned into training and testing sets.
4. **Model Evaluation**: K-fold cross-validation was performed on the training data. Five fundamental machine learning classifiers and our proposed ensemble models were thoroughly assessed to identify the best-performing model.
5. **Model Optimization**: The best model was further optimized using hyperparameter tuning and subsequently evaluated on the test data.
6. **User Interface Development**: A user-friendly interface was designed to enable seamless application of the liver disease prediction model, providing accessible and actionable insights.

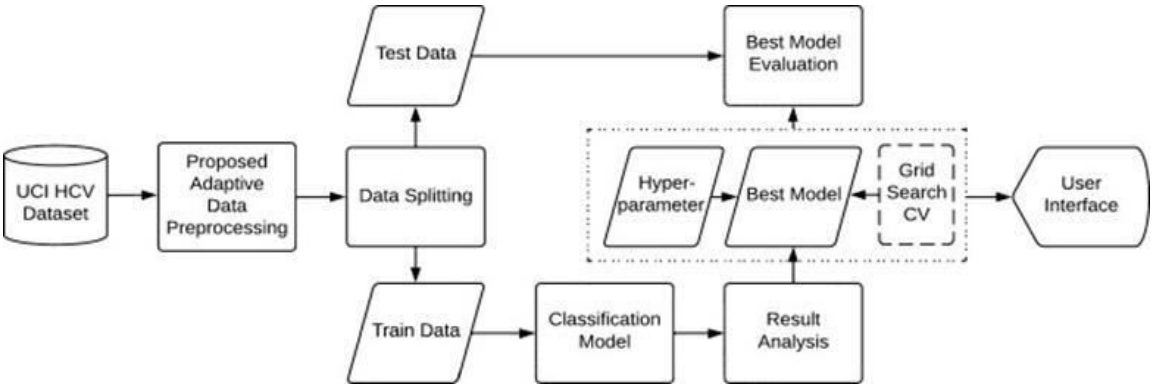


Fig 1. Block diagram of our proposed liver disease prediction system.

3.1 Dataset description

We conducted our analysis on a real-world dataset comprising 615 instances, containing clinical and laboratory measurements of patients. The dataset includes demographic information, blood test results, and other relevant medical features. It consists of five primary classes and 12 attributes, where all attributes, except for **Sex**, are numerical. Among these, 10 attributes represent laboratory data.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404382|

Table 1 provides a detailed description of these 10 attributes, along with Age and Sex, along with a brief statistical summary. Additional attribute information was gathered from relevant sources. Furthermore, Figure 2 presents histograms illustrating the distribution of the 10 attributes, as well as Age and Sex.

Table 1. Statistical summary of 12 attributes and their descriptions.

SN	Attributes	Description	Mean ± STD
1	AGE	Age (years)	47.41 ± 10.06
2	SEX	377 data from male and 238 data from female (male/female)	NAN
3	ALB (Albumin)	Protein synthesized by the liver, contributes to immune function and acts as a binding agent for certain molecules (g/L)	41.62 ± 5.78
4	ALP (Alkaline Phosphatase)	An enzyme, plays an integral role in metabolism within the liver and development within the skeleton (u/L)	68.28 ± 26.03
5	ALT (Alanine Aminotransferase)	An important enzyme in the intermediary metabolism of glucose and protein (u/L)	28.45 ± 25.47
6	AST (Aspartate Aminotransferase)	An enzyme in amino acid metabolism found in liver, heart, skeletal muscle, etc. (u/L)	34.79 ± 33.09
7	BIL (Bilirubin)	Yellow pigment produced during the breakdown of red blood cells (µmol/L)	11.40 ± 19.67
8	CHE (Cholinesterase)	An enzyme, involved in the breakdown of acetylcholine, a neurotransmitter (100u/L)	8.20 ± 2.21
9	CHOL (Cholesterol)	A fatty substance produced by the liver and obtained from dietary sources (mmol/L)	5.37 ± 1.13
10	CREA (Creatinine)	A waste product produced by muscles during normal metabolism (µmol/L)	81.29 ± 49.76
11	GGT (Gamma-Glutamyl Transferase)	Liver enzyme, responsible for absorption of amino acids from the glomerular filtrate and from the intestinal lumen (u/L)	39.53 ± 54.66
12	PROT (Protein)	An essential macronutrient involved in various physiological processes, including tissue repair and immune function (g/L)	72.04 ± 5.40

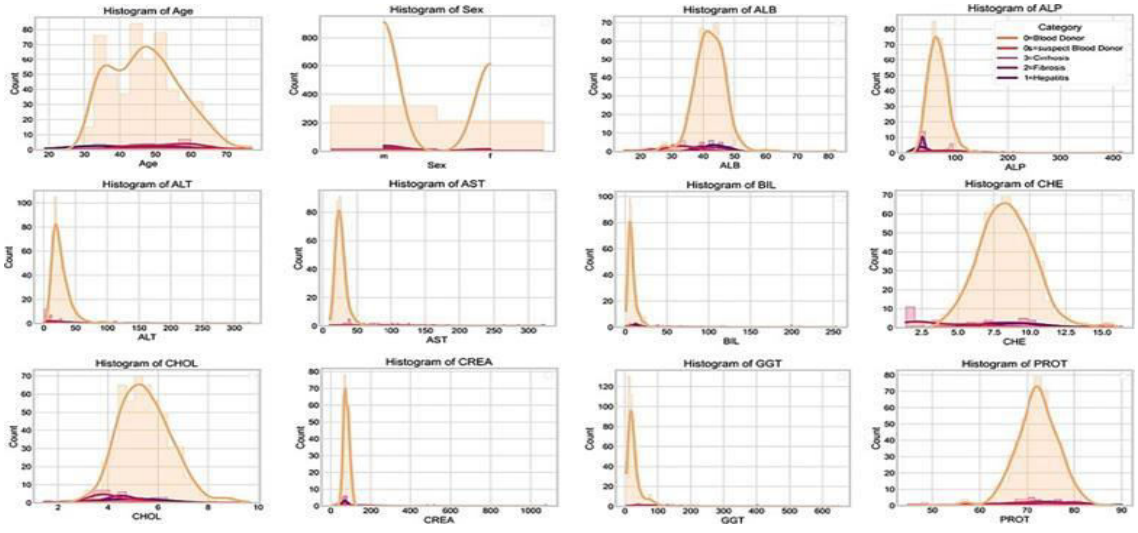


Fig 2. Histogram of all attributes including five categories.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404382|

3.2 Proposed adaptive data pre-processing

The dataset contained a total of 31 missing values across different attributes, distributed as follows: 1 in ALB, 18 in ALP, 1 in ALT, 10 in CHOL, and 1 in PROT.

Rather than applying a general imputation approach across all data, we adopted a class-specific mean imputation strategy. This decision was driven by the observation that the patterns of missing data exhibited distinct class-based characteristics. Using a unified imputation method would have compromised the class-specific information, reducing the discriminatory power of our model.

To address this, we calculated distinct mean values for each feature within its respective class. Each missing value in a feature was then replaced with the corresponding class-specific mean, ensuring that the imputed data retained its inherent class-based characteristics.

3.2.1 Outlier rejection

Outliers are the data points that have either extremely high or extremely low values that deviate significantly from the majority of the data and can distort our machine learning modeling [25]. We used the Interquartile Range (IQR) method. The IQR is defined as the range of values between the third, Q3 (75th percentile), and first, Q1 quartiles (25th percentile).

$$\text{LOWER Bound(LB)} = \text{Q1} - \text{k} \times \text{IQR} - - - (1)$$
$$\text{Upper Bound(UB)} = \text{Q3} - \text{k} \times \text{IQR} - - - - (2)$$

'k' is the scaling factor, and depending on the dataset, it can be modified. An outlier is defined as a data point that is below LB or greater than UB. After finding out the outlier, they are replaced with mean values. Our dataset displayed varying quantities of outliers across different categories shown in Table 2. We exclusively removed outliers from Category 0 (0=Blood Donor) due to its significant dominance, encompassing approximately 86.67% of the entire dataset. In contrast, the remaining categories primarily consisted of instances involving individuals with various medical conditions. This prompted us to refrain from conducting outlier removal procedures on these categories.

Table 2. Distribution of classes and outliers in the dataset.

Nominal value	Absolute count	Percentage	Outliers
0=Blood Donor	533	86.67%	125
0s=Suspected Blood Donor	7	1.14%	29
1=Hepatitis	24	3.90%	31
2=Fibrosis	21	3.41%	33
3=Cirrhosis	30	4.88%	100

3.2.2 Log normalization

Following outlier rejection, we applied log normalization to address skewed data, achieving a more uniform distribution by compressing larger values and stretching smaller ones. The goal was to make the dataset more symmetrical and suitable for our subsequent machine learning modeling.

3.2.3 Feature selection

The primary objective of the feature selection stage is to reduce the feature set to a smaller, more relevant subset to enhance classification accuracy. The SelectKBest method was used to identify the most significant features for the machine learning model. Since sex is a categorical variable, it was evaluated using a Chi-square test, while the remaining numerical features underwent an ANOVA F-test. ANOVA is a statistical technique that compares the means of different groups to derive the F-value, representing the ratio of variances between and within groups. A higher F-score indicates a stronger relationship between the feature and the target variable. On the other hand, the Chi-square test measures the association between categorical variables by comparing observed and expected frequencies. Figure 3 presents the results of both tests, with bar plots depicting the P-values and a line graph illustrating the F-scores. The analysis showed that 10 out of 12 features were statistically significant, with P-values below 5×10^{-9}. These features

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404382|

demonstrated strong discriminative power, as indicated by their high F-scores. Based on these results, the top 10 features were selected for model training and evaluation to ensure optimal predictive performance.

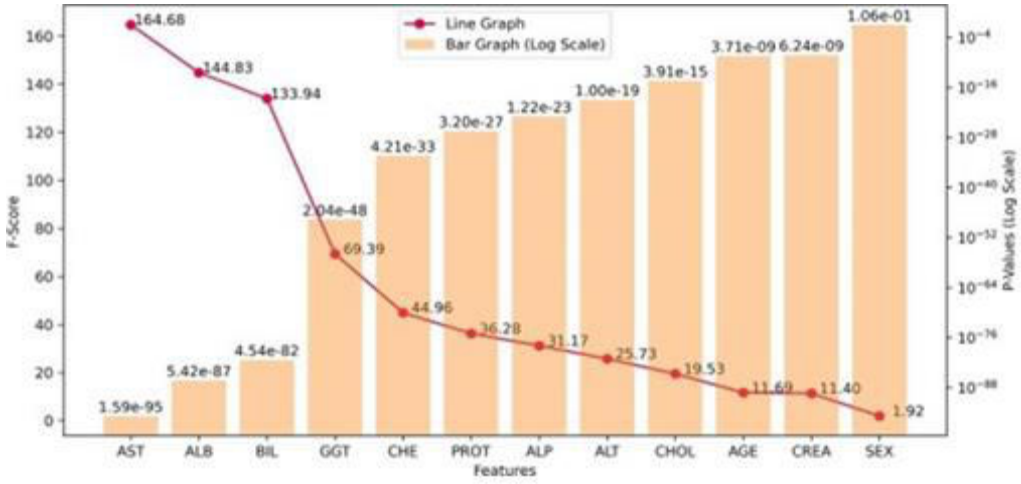


Fig. 3. Graph for F-score and P-Values using Anova F-test and Chi-square test respectively.

3.2.4 Feature scaling

We utilized StandardScaler for feature scaling, aiming to achieve a uniform scale. Employing the StandardScaler approach involved calculating the mean and standard deviation for each feature from the training data. The values were transformed to center around zero with a variance of 1 using the following formula:

$$z = \frac{x-u}{\sigma} - - - (3)$$

where x is the original value of the feature, μ is the mean of the feature, σ is the standard deviation of the feature. The application of StandardScaler is particularly beneficial for distance-based machine learning methods, such as LR and SVM, enhancing their performance by mitigating the impact of varying scales.

3.2.5 Data balancing

The initial dataset had an unbalanced distribution, with substantially more instances belonging to one class than the other. In order to mitigate this imbalance, SMOTE was employed to generate synthetic samples for the minority classes. The method strategically picks a minority class instance and its k nearest neighbors, then finds the difference between them which is multiplied by a constant, λ (0 <λ<1). Then it is added to the feature value of the instances [29]. New synthetic instances are generated using the following formula,

$$z = x + \lambda.(y - x) - - - (4)$$

where x represents a minority instance, y represents a nearest neighbor, and z represents a synthetic instance. Figure 4 shows the nested pie-chart before and after the data balancing stage. The outer circle shows the imbalanced data among the five categories and the inner circle shows the balanced data after involving SMOTE algorithm. It is evident from the chart that the SMOTE algorithm has generated additional instances for the imbalanced classes, resulting in uniform instances for all classes. This is depicted in the figure, where each class comprises 20% of the instances. The procedure for our suggested adaptive data preprocessing is outlined in Algorithm 1.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404382|

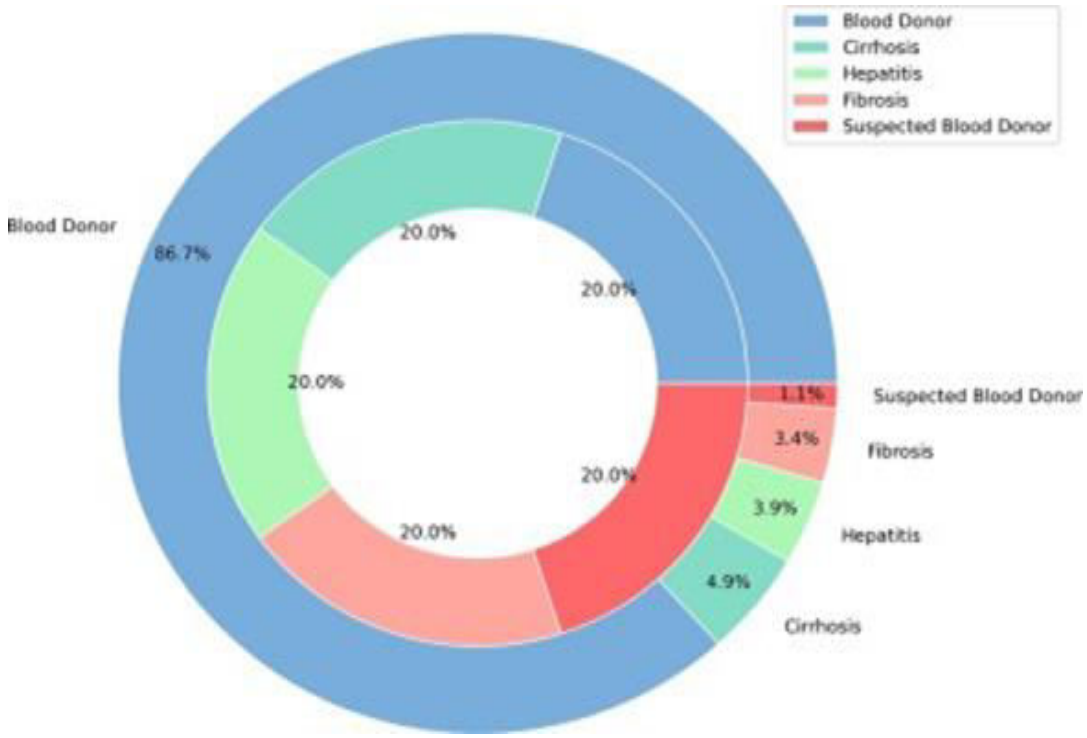


Fig. 4. Class distribution before and after applying SMOTE (inner circle representing after SMOTE and outer circle representing before SMOTE).

Input: Original Dataset $X \in \mathbb{R}^{r \times n}$, where r is the number of features and n is the number of samples.
Output: Balanced Data.
Step 1. Load Dataset
$X_{\text{input}} = \text{LoadInputData}$
Step 2. Class specific mean imputation
$X_{\text{filled}} = FillMissingValues(X_{\text{input}})$
Step 3. Remove outliers
$X_{\text{no\_outliers}} = RemoveOutliersFromCategory0(X_{\text{filled}})$
Step 4. Log Normalization
$\text{Skewed\_Features} = IdentifySkewedFeatures(X_{\text{no\_outliers}})$
for each feature **in** Skewed_Features **do**
  while *not EvenlyDistributed*$(X_{\text{no\_outliers}}[feature])$ **do**
    if *IsLogTransformable*$(X_{\text{no\_outliers}}[feature])$ **then**
      $X_{\text{normalized}}[\text{feature}] = LogTransform(X_{\text{no\_outliers}}[\text{feature}])$
    end
  end
end
Step 5. Feature Selection
$X_{\text{selected\_features}} = PerformFeatureSelection(X_{\text{normalized}})$
Step 6. Feature Scaling $X_{\text{scaled}} = StandardScaler(X_{\text{selected\_features}})$
Step 7. Data Balancing
$X_{\text{Balanced}} = SMOTE(X_{\text{scaled}})$

Algorithm 1. The Steps of Implementing Proposed Adaptive Data Preprocessing Technique.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404382|

3.3 Data splitting

Following SMOTE oversampling, each class had 533 instances. A distinct set of samples was reserved and kept separate from the training data. On the remaining training data, we performed K-fold cross-validation.

3.3.1 Test subset selection

In each class, a deliberate selection of 100 samples is made to create a distinct test subset using a stratified sampling approach. These examples were only used to judge how well the model did after it was trained.

3.3.2 K-fold cross-validation

K-fold cross-validation involves dividing the data set into k subsets. In each iteration, the model is trained using (k - 1) parts, while the remaining portion is allocated as a validation set. The number of folds determines how many times the process is to be iterated [30]. In our research study, we employed 5-fold cross-validation on the dataset.

3.4 Classification model

We utilized five distinct machine learning classifiers - LR, SVM, DT, RF, and XGBoost - to independently predict liver disease outcomes. Additionally, we detailed our proposed ensemble model combining these basic classifiers.

3.4.1 Logistic regression (LR)

LR is a supervised machine learning algorithm used for classification. For a given dataset, it assesses the relationship between the response (dependent) variable and one or more explanatory (independent) variables, indicating the importance and intensity of the explanatory variables' impact on the response variable. In contrast to multivariate linear regression, logistic regression differs in terms of the response variable and the type of analysis conducted [31]. The logistic function, commonly known as the sigmoid function, is typically used in LR to estimate probabilities [32]; it is given by,

$$p(x) = \frac{1}{1+e^{-(a+bx)}} - - - (5)$$

where P(x) is the probability, a is the slope, b is the y-axis intercept, and x is the independent variable. The term (a+bx) is derived from the equation of line. The logistic function is a simple S-shaped curve that converts data to a value between 0 and 1.

3.4.2 Support vector machine (SVM)

SVM [33] is a strong and probably the most widely used machine learning algorithm that has attracted considerable interest in various research fields. In our research study, non-linear SVM was chosen as a classifier because the data was not perfectly linearly separable. In cases when the data cannot be separated linearly, a kernel function is used to transform the data into a higher-dimensional space where it becomes linearly separable. RBF kernel, also known as Gaussian kernel, calculates how similar the two data points are based on the Euclidian distance between them with a free parameter σ. The kernel function is given by,

$$k(x, x') = \frac{\exp\left(-\left\|x-x'\right\|^2\right.}{2\sigma^2} - - - (6)$$

where $\|x-x'\|$ represents the Euclidean distance between x and x′, and σ2 is the variance. When x and x′ are very close (small $\|x-x'\|$), the exponent becomes close to 0, making K(x, x′) close to 1. As the distance between x and x′ increases, K(x, x′) decreases towards 0 [34].

3.4.3 Decision tree (DT)

DT is a supervised machine learning approach that may be used for both classification and regression analysis; however, it is most commonly employed for classification. It has a tree-like hierarchical structure with a root node, branches, internal nodes, and leaf nodes. Using the properties of the data as input, DT generates sequential, binary judgments at each node, thereby splitting the data at each node. These choices are based on the input feature values that minimize impurities and other variables to the greatest extent possible [35]. To implement it, the Gini Impurity or Information Gain is utilized as a metric to establish how the features of a dataset should divide nodes.
Information Gain=Parent entropy−Weighted sum of child entropy ---(7)

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404382|

In the information gain equation, parent entropy means the parent node before the split, and it is subtracted by the weighted sum of the child nodes after the split. Entropy is an idea that comes from the information theory which measures the uncertainty of sample values. The following equation provides its definition:

$$\text{Entropy} = -\sum_{i=1}^{N} p_i \log_2 p_i - - - (8)$$

where S represents the dataset, C represents the classes in set S and p(C) is the proportion of data points that belong to class C to the number of total data points in set S [36].

3.4.4 Random forest (RF)

RF is a type of ensemble learning algorithm built using DT algorithms [37]. While DT can be prone to problems, such as bias and overfitting, random forest comes with a better apprehending of its limitations while still maintaining its strength. It creates a 'forest' that is trained via bagging or bootstrap aggregation. Bagging involves creating several sample sets from a larger dataset using sampling with replacement. Some instances may appear numerous times and others may not appear at all due to the sampling with replacement, but the total size of each subset is the same as the original dataset. A separate DT is trained on each of these subsets. The final prediction is reached by taking the majority votes or the most frequent categorical variable from all of the trees [38].

3.4.5 XGBoost (XB)

eXtreme Gradient Boosting, abbreviated as XGBoost [39] is a DT-based ensemble machine learning algorithm that uses a gradient boosting framework. The name "gradient boosting" comes from the idea that a weak model can be strengthened by being combined with several additional weak models to produce a collectively strong one. It consists of additively generating weak models via a gradient descent algorithm connected to an objective function. In order to reduce errors as much as possible, gradient boosting sets goals for the next model. The desired result in each case is determined by the gradient of the error in relation to the prediction.

3.4.6 Proposed ensemble models

Stacking is an ensemble learning method that boosts predictive performance by integrating multiple models. In this approach, individual base models generate predictions, and these predictions serve as input features for a meta-model. The meta-model is then trained to learn how to weigh and combine these diverse predictions, often enhancing overall performance by capturing a broader range of patterns. Figure 5 depicts the architectural configuration of our proposed ensemble, incorporating the model stacking technique. In this ensemble framework, four base models are integrated alongside a final estimator. The predictions generated by the base models serve as the input for the final model, facilitating the ultimate prediction. Table 3 illustrates the various combinations of base models and final estimators employed in the creation of our proposed ensemble models.

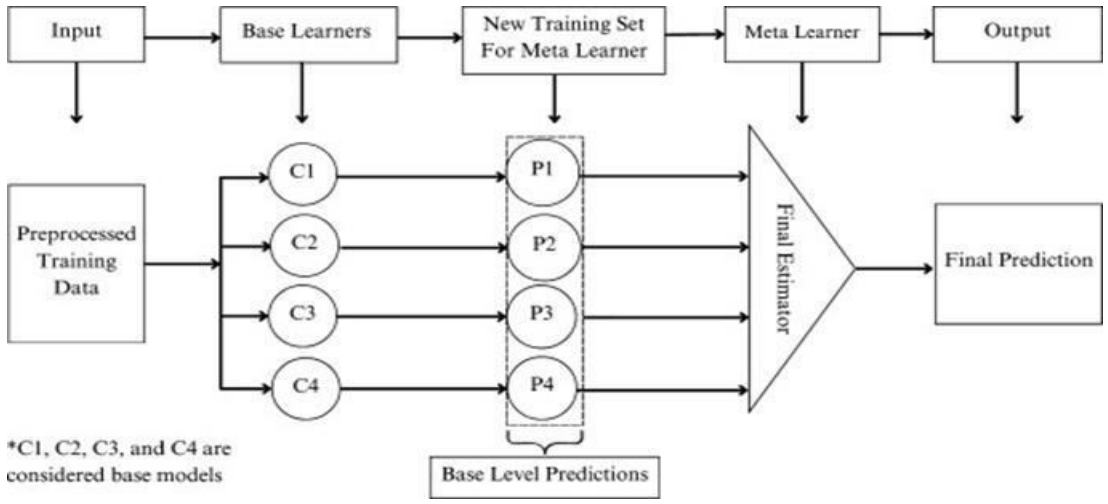


Fig. 5. Architectural structure of our proposed ensemble models.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404382|

Table 3. Base model and final estimator combinations in our proposed ensemble models.

Name	Base model	Final estimator
Ensemble LR	XB+RF+SVM+DT	LR
Ensemble XB	RF+SVM+DT+LR	XB
Ensemble RF	SVM+DT+LR+XB	RF

IV. RESULT ANALYSIS

4.1 Performance Metrics

The performance of the model is analyzed by using the confusion matrix. This will specify the performance of classification models for given test data. This will specify the values for test data that are known. This matrix is divided into two attributes such as predicted values and original values along with an overall number of predictions.

True Negative (TN): The prediction value is false and actual value is also false.
True Positive (TP): The prediction value is true and actual value is false.
False Positive (FP): The predicted value is true and actual value is false.
False Negative (FN): The predicted value is false and actual value is true.

The effectiveness of a machine learning model can be measured via evaluation metrics. These metrics offer valuable insights into the degree of alignment between a model's predictions and the actual outcomes, enabling an impartial evaluation of the model's quality and efficacy. Below is a quick summary of each of the five evaluation matrices that we measured :

Accuracy: One of the most popular evaluation matrices is accuracy. It is calculated as the percentage of predictions that the model gets correct. The accuracy is formulated as,

$$\text{Acuuracy} = \frac{\text{Number of correct predictions}}{\text{Total Number of predictions}} - - - (9)$$
$$= \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} - - - (10)$$ -

where TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative.

Macro Average Precision (MAP): Precision refers to the proportion of accurately predicted positive cases out of all instances predicted as positive. In multi-class classification, it is possible to compute precision for each individual class in an independent manner. To ensure that all classes are given the same weight, MAP takes the mean of all precision values. The MAP is formulated as,

$$\text{precision} = \frac{\text{TP}}{\text{TP+FP}} - - - (11)$$
$$\text{MAP} = \frac{\sum_{i=1}^{n} \text{precision}}{\text{N}} - - - (12)$$

where N is the total number of classes.

Macro Average Recall (MAR): Recall calculates the percentage of accurately predicted positive instances out of all actual positive instances. It is also known as sensitivity or true positive rate.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FP}} - - - (13)$$
$$\text{MAR} = \frac{\sum_{i=1}^{n} \text{Recall}}{\text{N}} - - - (14)$$

Mean Absolute Error (MAE): It is defined as the average absolute difference between the predicted values of the model and the true values of the data. MAE provides a straightforward depiction of the model's accuracy in absolute errors, with lower numbers indicating greater performance. The MAE is simple to understand and is well-suited for

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404382|

cases where the goal is to minimize the model's average error. It is frequently used alongside other evaluation metrics such as the root mean squared error (RMSE). It is formulated as,

$$\text{MAE} = \frac{\sum_{i=1}^{N} |y_i - y_i'|}{N} - - - (15)$$

in which, N is the total number of classes, y is the actual value and y′ is the predicted value.

Root Mean Squared Error (RMSE): The root mean square error (RMSE) is computed by taking the square root of the mean of the squared differences between the predicted values and the actual values. RMSE gives more weight to larger errors, making it suitable when larger errors are of greater concern compared to MAE. It is formulated as

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (y_i - y_i)^2}{N}} - - - (16)$$

in which, N is the total number of classes, y is the actual value and y′ is the predicted value.
This section presents the effect of our adaptive data preprocessing technique with the corresponding results in several subsections. We detailed the preprocessing findings and the evaluation of different ML models.

4.2 Effect of outlier rejection

As discussed earlier, it can be more challenging for the model to understand the underlying patterns when outliers distort the data distribution. Hence, we intended to remove the outliers for a more accurate representation of the data. To illustrate the impact of outlier rejection, we utilized boxplots to visualize the data distribution before and after applying the IQR method Figure 6.

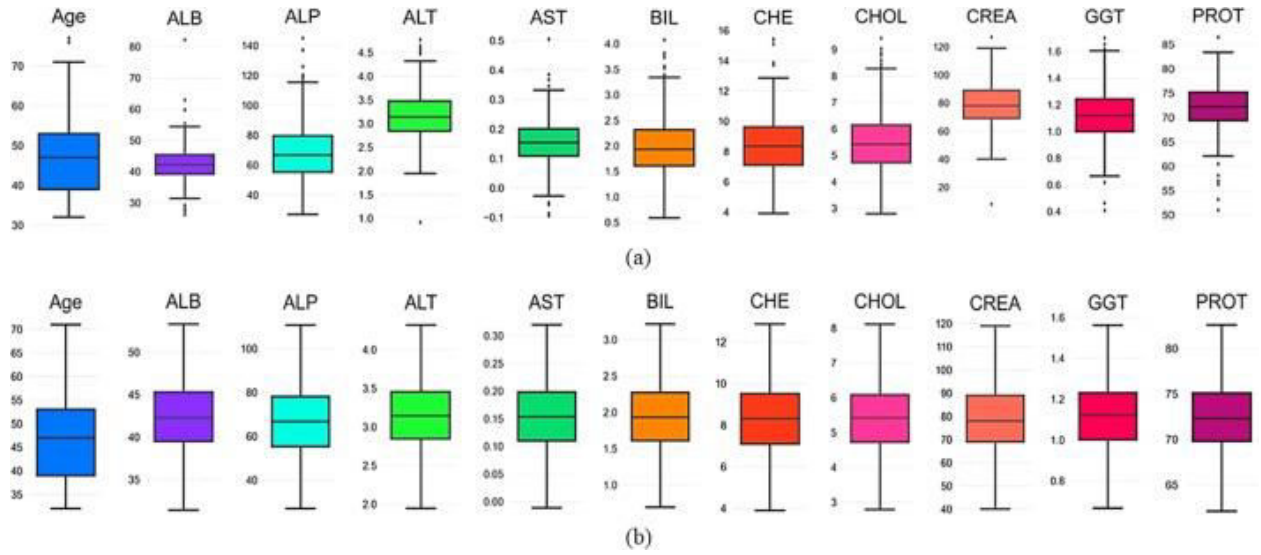


Fig. 6. The distribution of attributes using box plot (a) before and (b) after outlier removal for Category 0.

4.3. Effect of log normalization

The pre-normalization QQ plot, shown in Figure 7, revealed a notable deviation from the diagonal line, suggesting major deviations from a normal distribution and the presence of a heavy tail in the dataset. A distribution that is heavily skewed has the potential to introduce bias into model predictions and result in inferior performance. Hence, following the rejection of outliers, log transformation was conducted to mitigate the skewness observed in the data. It also scaled down large values to a more manageable level. Figure 7, Q-Q plot, shows the effect of log normalization for these 4 features. For highly skewed data e.g., AST and GGT, single log transformation was not enough to fully address the skewness. We used double and triple log normalization for GGT and AST respectively for more amenable analysis.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404382|

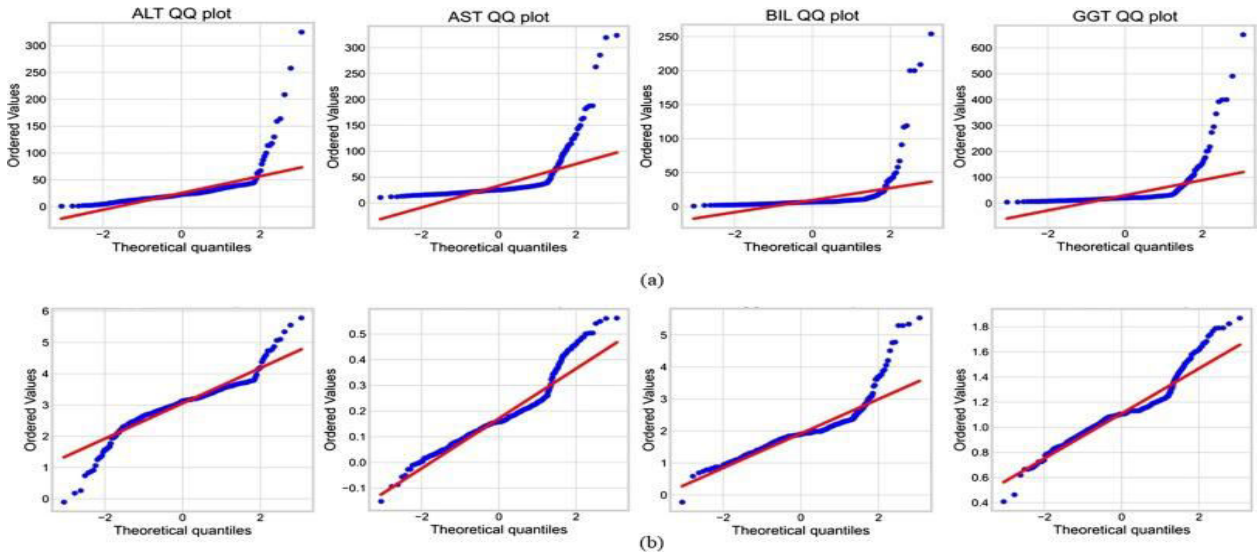


Fig. 7. QQ plot (a) before and (b) after log transformation.

4.4. Evaluation of the proposed data preprocessing technique without data balancing

Table 4 presents the outcomes, demonstrating that our innovative approach involving mean imputation, outlier rejection, and log transformation, as detailed in Section 4, significantly enhances classifier accuracy. To assess the efficacy of our proposed method, we conducted a comparative analysis with the commonly used conventional standard data preprocessing technique in machine learning. In the case of our proposed data preprocessing method without data balancing, we implemented all steps outlined in our approach, excluding the specific data balancing step. In the standard data processing case, we employed standard mean imputation to fill missing values, followed by the identical feature selection and scaling steps as detailed in our proposed data processing technique. Subsequently, both scenarios underwent k-fold cross-validation to assess the performance of our various classifiers. Particularly noteworthy were the Ensemble RF and Ensemble XB models, showing remarkable gains in accuracy at 96.419% and 96.257%, respectively. Furthermore, there was a significant reduction in MAE and RMSE values, indicative of decreased prediction errors. Among the five basic ML classifiers, LR demonstrated a substantial increase in accuracy, reaching 96.256%, accompanied by noteworthy enhancements in precision MAP and MAR.

Table 4. Comparison of classifier performance with the proposed and conventional preprocessing methods.

Type of data preprocessing	Classifier's name	Accuracy (%)	MAP	MAR	MAE	RMSE
Standard Data Pre-processing	LR	93.001	0.8253	0.6066	0.1432	0.6064
	DT	90.782	0.6502	0.5234	0.1887	0.7124
	RF	92.034	0.7583	0.4984	0.1409	0.5634
	SVM	91.700	0.7829	0.4416	0.1839	0.6998
	XB	92.846	0.7224	0.5336	0.1253	0.5401
	Ensemble LR	93.167	0.7132	0.6436	0.1205	0.5008
	Ensemble XB	93.980	0.7183	0.6966	0.0976	0.4990
	Ensemble RF	93.333	0.6972	0.5751	0.1205	0.5307
Proposed Data preprocessing without Data balancing	LR	96.256	0.9034	0.7698	0.0686	0.4133
	DT	92.687	0.6626	0.6052	0.1364	0.5713
	RF	95.171	0.7884	0.7272	0.0767	0.4417
	SVM	95.416	0.8434	0.6255	0.0797	0.4087
	XB	94.964	0.7887	0.6849	0.0927	0.4739
	Ensemble LR	95.281	0.7640	0.6829	0.0798	0.4290
	Ensemble XB	96.257	0.8159	0.7758	0.0489	0.2823
	Ensemble RF	96.419	0.8275	0.7881	0.0423	0.2707

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404382|

V. CONCLUSION

In our Project, we explored liver disease prediction through a machine learning approach utilizing the UCI HCV dataset. Our proposed adaptive data preprocessing technique significantly enhanced classifier performance compared to the standard data preprocessing technique across various metrics. We utilized five fundamental machine learning algorithms and combined them to create three additional ensemble models, further improving the performance of liver disease prediction. Following extensive analysis, the selected best model, the hyper-tuned ensemble LR, achieved an impressive accuracy of 99.80% on the testing data and obtained a perfect AUC score of 1. This superior performance demonstrates that our liver disease prediction model surpasses previous research efforts.

Our collaborative research effort has significantly contributed to advancing HCV prediction through a comprehensive methodology that includes dataset exploration, innovative data preprocessing technique, fine-tuning, and user interface development. Our study can help improve healthcare decision-making by combining scientific rigor and practicality, as well as highlighting machine learning's novel capabilities in improving medical diagnostics and patient care.

REFERENCES

[1] R. Bhardwaj, A. R. Nambiar, and D. Dutta, "A study of machine learning in healthcare," in *Proc. 41st Annu. Comput. Softw. Appl. Conf. (COMPSAC)*, vol. 2, IEEE, 2017, pp. 236–241.
[2] M. M. Ahsan, S. A. Luna, and Z. Siddique, "Machine-learning-based disease diagnosis: a comprehensive review," *Healthcare*, vol. 10, p. 541, 2022.
[3] S. K. Asrani, H. Devarbhavi, J. Eaton, and P. S. Kamath, "Burden of liver diseases in the world," *J. Hepatol.*, vol. 70, no. 1, pp. 151–171, 2019.
[4] A. M. Moon, A. G. Singal, and E. B. Tapper, "Contemporary epidemiology of chronic liver disease and cirrhosis," *Clin. Gastroenterol. Hepatol.*, vol. 18, no. 12, pp. 2650–2666, 2020.
[5] D. Iluz-Freundlich, M. Zhang, J. Uhanova, and G. Y. Minuk, "The relative expression of hepatocellular and cholestatic liver enzymes in adult patients with liver disease," *Ann. Hepatol.*, vol. 19, no. 2, pp. 204–208, 2020.
[6] V. Lala, M. Zubair, and D. A. Minter, "Liver function tests," *StatPearls* [Internet], StatPearls Publishing, 2022.
[7] S. Blach et al., "Global change in hepatitis C virus prevalence and cascade of care between 2015 and 2020: a modelling study," *Lancet Gastroenterol. Hepatol.*, vol. 7, no. 5, pp. 396–415, 2022.
[8] D. Schuppan and Y. O. Kim, "Evolving therapies for liver fibrosis," *J. Clin. Invest.*, vol. 123, no. 5, pp. 1887–1901, 2013.
[9] M. Pinzani, M. Rosselli, and M. Zuckermann, "Liver cirrhosis," *Best Pract. Res. Clin. Gastroenterol.*, vol. 25, no. 2, pp. 281–290, 2011.
[10] H. M. 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," *Knowl. Inf. Syst.*, vol. 65, no. 6, pp. 2595–2617, 2023.
[11] K. Ahammed, M. S. Satu, M. I. Khan, and M. Whaiduzzaman, "Predicting infectious state of hepatitis C virus affected patient's applying machine learning methods," in *Proc. IEEE Reg. 10 Symp. (TENSYMP)*, IEEE, 2020, pp. 1371–1374.
[12] H. Ayeldeen, O. Shaker, G. Ayeldeen, and K. M. Anwar, "Prediction of liver fibrosis stages by machine learning model: a decision tree approach," in *Proc. 3rd World Conf. Complex Syst. (WCCS)*, IEEE, 2015, pp. 1–6.
[13] R. Amin, R. Yasmin, S. Ruhi, M. H. Rahman, and M. S. Reza, "Prediction of chronic liver disease patients using integrated projection based statistical feature extraction with machine learning algorithms," *Inform. Med. Unlocked*, vol. 36, Art. no. 101155, 2023.
[14] S. Islam, A. U. Rehman, S. Javaid, T. M. Ali, and A. Nawaz, "An integrated machine learning framework for classification of cirrhosis, fibrosis, and hepatitis," in *Proc. 3rd Int. Conf. Latest Trends Elect. Eng. Comput. Technol. (INTELLECT)*, IEEE, 2022, pp. 1–6.
[15] P. Verma, K. Deshmukh, and D. Krishnan, "Multiclass classification of liver diseases using optimized machine learning classifiers," in *Proc. Int. Conf. Comput. Intell. Knowl. Economy (ICCIKE)*, IEEE, 2023, pp. 461–466.
[16] F. Mostafa, E. Hasan, M. Williamson, and H. Khan, "Statistical machine learning approaches to liver disease prediction," *Livers*, vol. 1, no. 4, pp. 294–312, 2021.
[17] 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.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404382|

[18] A. Sivasangari, B. J. K. Reddy, A. Kiran, and P. Ajitha, "Diagnosis of liver disease using machine learning models," in *Proc. 4th Int. Conf. I-SMAC (IoT Soc. Mobile Analytics Cloud)*, IEEE, 2020, pp. 627–630.
[19] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. Quinn, and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: performance analysis and comparison," *Comput. Biol. Med.*, vol. 136, Art. no. 104672, 2021.
[20] K. F. Lichtinghagen and R. G. Hoffmann, "HCV data, UCI machine learning repository," 2020. [Online]. Available: https://doi.org/10.24432/C5D612.
[21] 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, no. 6, 2018.
[22] Y. Huang, D. Yang, and X. Zhang, "Prediction of hepatitis C based on liver function test features," in Proc. Int. Conf. Stat., Appl. Math., and Comput. Sci. (CSAMCS), vol. 12163, SPIE, 2022, pp. 429–435.
[23] M. L. Bishop, E. P. Fody, and L. E. Schoeff, Clinical Chemistry: Techniques, Principles, and Correlations, Alphen Aan Den Rijn, The Netherlands: Wolters Kluwer, 2018.
[24] N. Rifai, Tietz Textbook of Clinical Chemistry and Molecular Diagnostics, 6th ed., Elsevier Health Sciences, 2017.
[25] G. Barbato, E. Barini, G. Genta, and R. Levi, "Features and performance of some outlier detection methods," J. Appl. Stat., vol. 38, no. 10, pp. 2133–2149, 2011.
[26] H. Vinutha, B. Poornima, and B. Sagar, "Detection of outliers using interquartile range technique from intrusion dataset," in Proc. 6th Int. Conf. FICTA, Springer, 2018, pp. 511–518.
[27] N. O. F. Elssied, O. Ibrahim, and A. H. Osman, "A novel feature selection based on one-way anova f-test for e-mail spam classification," Res. J. Appl. Sci., Eng. Technol., vol. 7, no. 3, pp. 625–638, 2014.
[28] A. Agresti, Categorical Data Analysis, 2nd ed., vol. 792, John Wiley & Sons, 2012.
[29] 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.
[30] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," Stat. Surv., 2010.
[31] D. W. Hosmer Jr., S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3rd ed., vol. 398, John Wiley & Sons, 2013.
[32] J. Shu et al., "Clear cell renal cell carcinoma: CT-based radiomics features for the prediction of Fuhrman grade," Eur. J. Radiol., vol. 109, pp. 8–12, 2018.
[33] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," IEEE Intell. Syst. Appl., vol. 13, no. 4, pp. 18–28, 1998.
[34] S. Ghosh, A. Dasgupta, and A. Swetapadma, "A study on support vector machine based linear and non-linear pattern classification," in Proc. Int. Conf. Intell. Sustain. Syst. (ICISS), IEEE, 2019, pp. 24–28.
[35] S. Suthaharan, "Decision tree learning," in Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning, Springer, 2016, pp. 237–269.
[36] A. S. Karthiga, M. S. Mary, and M. Yogasini, "Early prediction of heart disease using decision tree algorithm," Int. J. Adv. Res. Basic Eng. Sci. Technol., vol. 3, no. 3, pp. 1–17, 2017.
[37] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining, Boston: Pearson Education, 2006.
[38] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning, vol. 112, Springer, 2013.
[39] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016, pp. 785–794.
[40] S. Raschka, "An overview of general performance metrics of binary classifier systems," arXiv preprint, arXiv:1410.5330, 2014.
[41] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," Int. J. Data Min. Knowl. Manag. Process, vol. 5, no. 2, p. 1, 2015.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details