SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V12I2P101

Volume 12 Issue 2, 1-11, February 2025

Original Article

Robust and Explainable Ensemble Based Framework for Liver Disease Classification using Data Balancing and Upsampling

Aditya Bhongade[1], Yogita Dubey[2], Prachi Palsodkar[3], Punit Fulzele[4]

[1,2]*Department of Electronics and Telecommunication Engineering, Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India.*
[3]*Department of Electronics Engineering, Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India.*
[4]*Directorate of Research and Innovation, SPDC, Datta Meghe Institute of Higher Education & Research, Wardha, Maharashtra, India.*

[2]*Corresponding Author : yogeetakdubey@yahoo.co.in*

Received: 04 December 2024 Revised: 03 January 2025 Accepted: 02 February 2025 Published: 26 February 2025

Abstract *- Liver Disease (LD) is a severe health condition impacting over 2 million lives annually worldwide, as reported by the WHO. Factors such as rising alcohol consumption, increased type 2 diabetes cases, genetic predispositions, and various lifestyle influences are expected to heighten LD prevalence further, underscoring the need for a modern, accurate, and interpretable classification system. This paper aims to develop an effective and transparent Machine Learning (ML) framework using ensemble learning models and Explainable AI (XAI) techniques for LD classification. The proposed framework addresses dataset imbalance and size constraints by employing data balancing and upsampling, enabling the ensemble models to learn complex patterns in clinical data. The performance of each model is evaluated, and the best-performing model, Gradient Boosting (GB), is further analyzed using SHAP, LIME, and ELI5 to interpret its feature impact. GB achieved high classification metrics, including accuracy, precision, recall, specificity, and AUC, with Direct Bilirubin, Alkaline Phosphatase, Alanine Aminotransferase, and Age identified as key influential features. This paper successfully presents a reliable and interpretable ML-based framework for LD classification, combining quantitative performance and explainability, making it highly suitable for clinical application.*

Keywords *- Classification, Ensemble learning, Explainable model, Feature analysis, Liver Disease.*

1. Introduction

Liver diseases have become a significant global health concern, with traditional diagnostic methods often proving inefficient and lacking in accuracy. In response, Machine Learning (ML) technologies have emerged as powerful tools for enhancing liver disease prediction, diagnosis, and treatment. Recent algorithms and computational power advancements have driven considerable interest in machine learning-based liver disease prediction. Numerous studies have explored various ML approaches aimed at improving the accuracy and consistency of disease classification and prediction, making them a promising solution in the medical field. Every year, around 2 million individuals die from liver disease throughout the world. According to the "Global Burden of Disease" research published in BMC Medicine, one in every four fatalities from cirrhosis and approximately one million deaths from liver cancer occurred in 2010 [1]. Liver disease consists of chronic disorders that affect liver function, frequently going through four stages: hepatitis, fibrosis, cirrhosis, and, finally, liver failure. Hepatitis, which is typically caused by viral infections or alcohol, causes inflammation in the liver.

Chronic inflammation, if left untreated, leads to fibrosis, which is scar tissue formation that impairs liver regeneration. Continued damage leads to cirrhosis, which is characterized by extensive scarring and reduced liver function that is generally permanent. Finally, liver failure is the end stage, in which the liver can no longer support crucial processes, needing urgent measures such as transplantation. Early identification and therapy are critical for slowing or halting this development [2].

Effective ML algorithms like Logistic Regression (LR), K-Nearest Neighbour (KNN), Decision Tree (DT), Support Vector Machine (SVM), ensemble algorithms including Extra Trees (ET), Random Forest (RF), AdaBoost, XGBoost, and Gradient Boosting (GB), along with feature scaling are used for early-stage detection of liver diseases [3-10].

Explainable AI (XAI) is the transparent AI in healthcare, showing that explainable predictions build trust and provide clear insights for both medical professionals and patients [11]. These methodologies and insights have broader applications across healthcare, emphasizing the need for accuracy and explainability in improving diagnosis and treatment strategies. In [12], RF, XGBoost, and Explainable Boosting Machine (EBM) achieved the highest accuracy at 99.8%, proving its superior effectiveness in liver disease prediction. XAI approaches are critical in solving the difficulty of transparency in AI models because they provide explicit and understandable explanations of how algorithms create certain outputs. Several strategies, including Local Interpretable Model-Agnostic Explanations (LIME), ELI5, Accumulated Local Effects (ALE) and Shapley Additive Explanations (SHAP), have received substantial attention in the literature. These methodologies provide more in-depth knowledge of model behaviour, improving the interpretability and dependability of AI systems, especially in vital domains such as the medical field and healthcare, where transparent decision-making is essential [13].

2. Related Work

A number of classification methods have been used in the literature to predict liver illness, including LR, KNN, and SVM. Furthermore, ensemble tree-based algorithms, including DT, RF, ET, AdaBoost, and XGBoost, are commonly employed. To improve the interpretability of these predictive models, XAI approaches such as ALE SHAP, ELI5, and LIME have been used, providing greater insights into how predictions are formed and increasing transparency and trust in model outputs, particularly in healthcare applications.

Afreen et al. (2021) introduced a novel boosting technique for liver disease classification, showing how ensemble learning enhances prediction accuracy [3]. Similarly, Shobana and Umamaheswari (2021) employed gradient boosting with feature scaling, underscoring the importance of preprocessing in improving classification. Singh et al. (2021) conducted a comparative analysis of various ML models, highlighting the effectiveness of DT and SVM [4]. Singh et al. (2021) carried out a performance analysis of ML algorithms for liver disease classification [5].

Dutta et al. (2022) focused on early-stage detection of liver diseases using machine learning algorithms, stressing the potential for early interventions [6]. Sokoliuk et al. (2020) explored binary classification approaches, proving the applicability of machine learning in differentiating between diseased and healthy cases [7]. Comparative analyses of algorithms by Ghosh et al. (2021) and Rabbi et al. (2020) reaffirmed the dominance of ensemble learning techniques in liver disease prediction, while Gupta et al. (2022) explored classification models for liver disease, demonstrating high accuracy rates. More recent studies have integrated explainable AI to interpret biomarkers associated with liver conditions, as evidenced by Arya et al. (2023) and Nilofer and Sasikala (2023), advancing transparency and trust in AI-driven medical predictions [8-12].

In specialized liver conditions like cirrhosis and Hepatitis C, explainable ensemble models were effectively employed by Alotaibi et al. (2023). Pei et al. (2021) also applied ML techniques to predict fatty liver disease, further broadening the scope of ML applications in liver health monitoring [12, 13]. Jei et al. (2024) created an explainable ML model for predicting High-Risk Nonalcoholic Steatohepatitis (NASH), attaining high accuracy and providing interpretable findings, which is critical for clinical decision-making and increasing trust in AI outputs [15]. Table 1 summarizes the existing work for liver disease classification in terms of methods used and key findings.

Table 1. Existing work for Liver Disease classification with the key finding

Literature	Method Used	Key Finding
Afreen et al. [3]	Boosting Algorithm	Improved Classification
Shobana & Umamaheswari [4]	Gradient Boosting with Feature Scaling	Enhanced Prediction
Singh et al. [5]	Performance Analysis of Algorithms	Algorithm Comparison
Dutta et al. [6]	Machine Learning Algorithms	Early Detection
Sokoliuk et al. [7]	Binary Classification Algorithms	Effective Classification
Ghosh et al. [8]	Comparative Analysis of Algorithms	Predictive Performance
Rabbi et al. [9]	Comparative Study of Algorithms	Model Effectiveness
Gupta et al. [10]	Classification Techniques	Disease Prediction
Arya et al. [11]	Explainable AI Techniques	Biomarker Interpretation
Nilofer & Sasikala et al. [12]	Comparative Study Using Explainable AI	Enhanced Interpretability
Alotaibi et al. [13]	Explainable Ensemble Models	Detecting Cirrhosis
Pei et al. [14]	Machine Learning Algorithms	Fatty Liver Prediction
Njei et al. [15]	Explainable Machine Learning Model	High-Risk Prediction

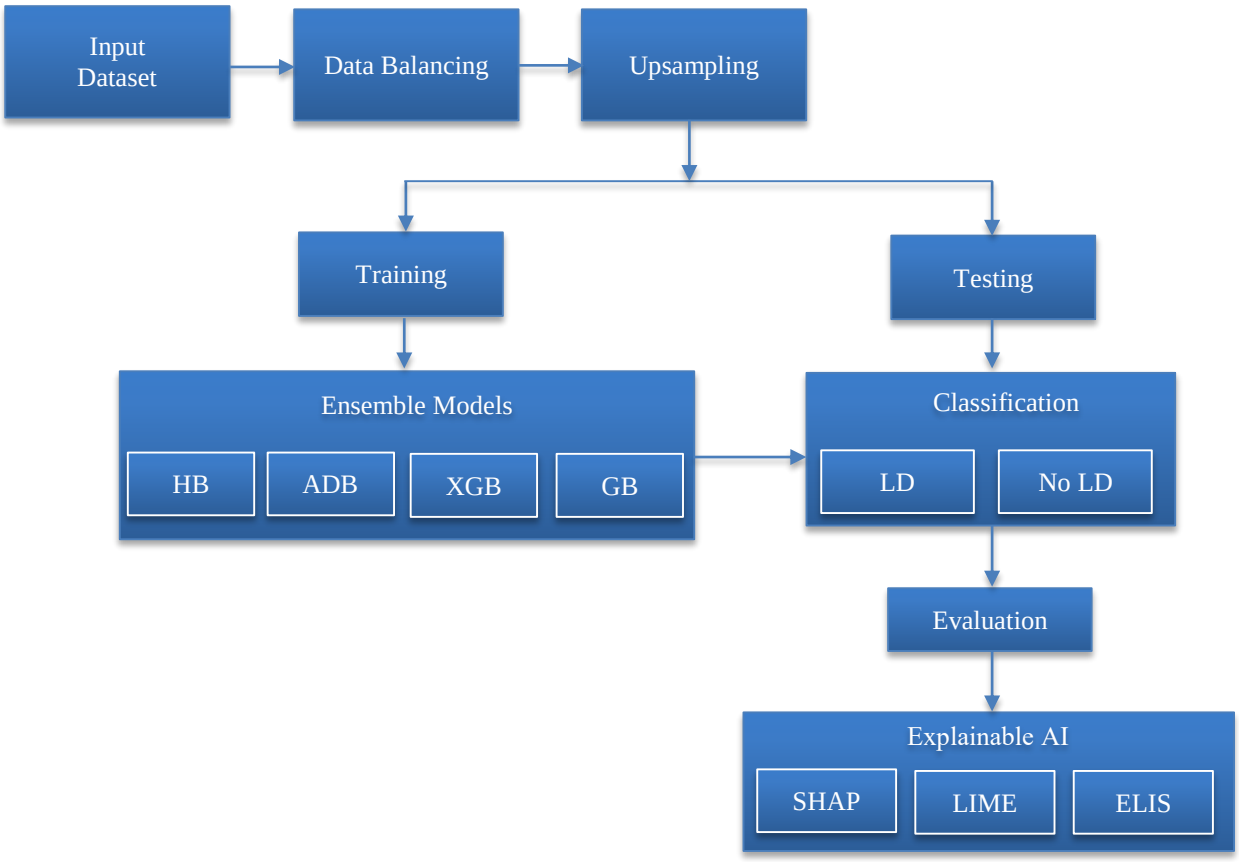

Fig. 1 Block diagram of proposed framework for Liver Disease classification with explainable model

Collectively, these studies indicate the critical role machine learning plays in liver disease prediction, enabling early detection and informed medical decisions. Emerging trends, such as the incorporation of explainable AI with boosting algorithms, resampling, use of tuned hyperparameters [16], folding, and feature scaling, are enhancing the interpretability of these models, making them more viable for clinical application [17-22].

3. Materials and Methods

The complete workflow of the proposed method for Liver disease classification is illustrated in Figure 1. The classification task starts with acquiring the dataset and ends with explaining the trained models using explainable AI agents.

3.1. Dataset Description

The dataset is sourced from publicly available medical data of liver patients from Andhra Pradesh, India [23]. It contains data from 583 subjects, out of which 416 have liver disease, and 167 do not. In this dataset, there are 10 input features and a bi-class target variable, i.e., Liver Disease or No Liver Disease. Out of the 10 input features, 9 are numerical: "Age, Total Bilirubin, Direct Bilirubin, Alkaline Phosphatase, Alanine Aminotransferase, Aspartate Aminotransferase, Total Proteins, Albumin, and Albumin to Globulin Ratio".

Gender is the only categorical feature in this dataset, having two categories: Male and Female. The details of the numerical and categorical features are displayed in Table 2, along with their range, mean value, and correlation with liver disease diagnosis. It can be observed that age ranges from 4 to 90 years, with the average age being 44.75 years. It must be noted that Ages above 89 years are all assigned 90 by the data publishers. Age has a correlation of 0.13 with the diagnosis. Total Bilirubin ranges from 0.4 to 75.0, having a mean value of 3.30 and a correlation value of 0.22 with the diagnosis. Direct Bilirubin ranges from 0.1 to 19.7, with a mean of 1.49 and a correlation of 0.25 with the diagnosis.

Enzyme features alkaline phosphotase, alamine aminotransferase, and aspartate aminotransferase, ranging from 63.0 to 2110.0, 10.0 to 2000.0 and 10.0 to 4929.0, respectively, with an average value of 290.58, 80.71, 109.91 respectively. These enzymes have correlation values of 0.18, 0.16 and 0.15, respectively. Total proteins have a range of 2.7 to 9.6, with an average of 6.48 and a correlation of -0.03 with liver diagnosis. Albumin and Albumin to Globulin Ratios have the range of 0.9 to 5.5 and 0.3 to 2.8, with average values of 3.14 and 0.95, respectively. Albumin and Albumin-to-globulin ratios both have a correlation value of -0.16 with liver diagnosis. Gender as male or female has a very low correlation value of -0.08 with liver diagnosis.

Table 2. Feature values with range, mean and correlation with Liver Disease diagnosis

SN	Feature	Range	Mean	Correlation with Diagnosis
1.	Age	4 - 90	44.75	0.13
2.	Total Bilirubin	0.4 - 75.0	3.30	0.22
3.	Direct Bilirubin	0.1 - 19.7	1.49	0.25
4.	Alkaline Phosphotase	63.0 - 2110.0	290.58	0.18
5.	Alamine Aminotransferase	10.0 - 2000.0	80.71	0.16
6.	Aspartate Aminotransferase	10.0 - 4929.0	109.91	0.15
7.	Total Protiens	2.7 - 9.6	6.48	-0.03
8.	Albumin	0.9 - 5.5	3.14	-0.16
9.	Albumin to Globulin Ratio	0.3 - 2.8	0.95	-0.16
10.	Gender	Male or Female		-0.08

3.2. Feature Analysis

An exhaustive feature analysis was performed on the dataset to thoroughly understand the input features. This section provides a detailed description of the feature analysis process, accompanied by intuitive figures. Figure 2 shows the count of Liver Disease (LD) and No LD cases for both genders, male and female.

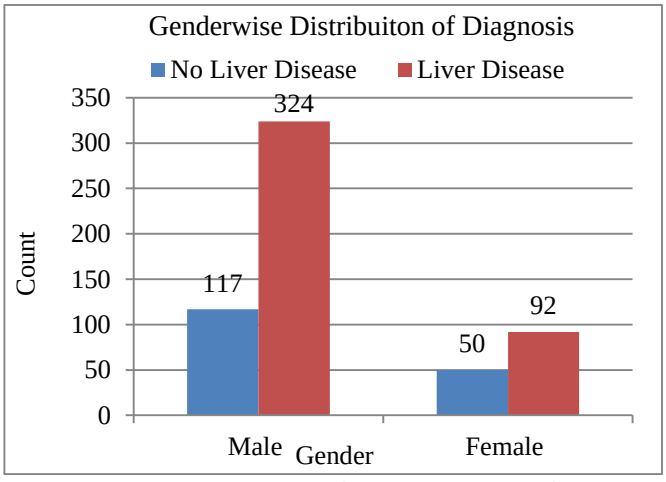

Fig. 2 Gender-wise distribution for Liver Disease classification

The dataset includes 324 LD and 117 No LD cases among males and 92 LD and 50 No LD cases among females. The percentage of LD cases is 73.5% for males (100 * 324 / (324 + 117)) and 64.8% for females (100 * 92 / (92 + 50)). This minor difference suggests that the prevalence of LD is nearly equal across genders. Additionally, gender correlation is -0.08 with the diagnosis, indicating minimal association.

Figure 3 illustrates the maximum and minimum values, along with the interquartile range of the features of total bilirubin and direct bilirubin for both diagnosis classes. The figure shows that the 75th, 50th (median), and 100th percentiles of both total and direct bilirubin are higher for the LD classes. Additionally, total bilirubin and direct bilirubin have positive correlation values of 0.22 and 0.25, respectively, as per Table 2, with a positive LD diagnosis. This indicates that higher levels of total and direct bilirubin contribute to an increased likelihood of the output being 1 (LD). Moreover, total bilirubin and direct bilirubin have relatively high correlation values compared to other features, thus significantly impacting the model's output.

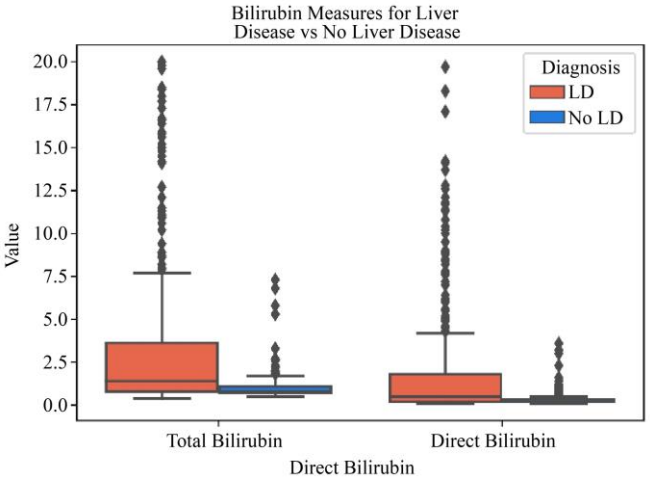

Fig. 3 Bilirubin measures for Liver Disease classification

Figure 4 illustrates the minimum, maximum, and interquartile ranges of the enzyme features alkaline phosphatase, alanine aminotransferase, and aspartate aminotransferase.

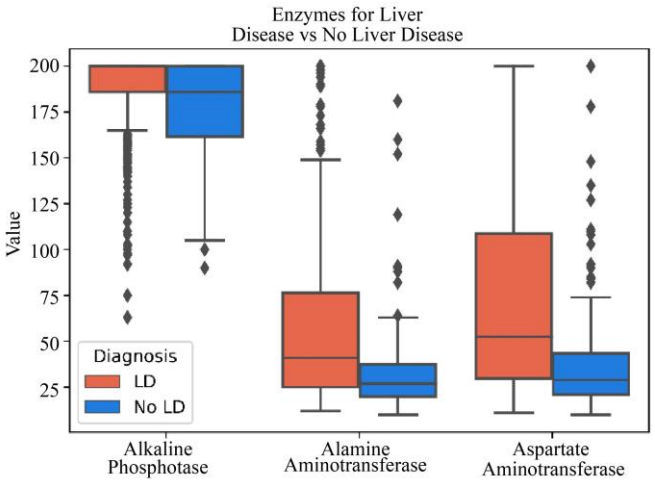

Fig. 4 Enzymes for Liver Disease classification

It can be observed that the 25th, 50th (median), 75th, and 100th percentile values of these enzyme features are higher for LD cases compared to No LD cases. The correlation values

for alkaline phosphatase, alanine aminotransferase, and aspartate aminotransferase are 0.18, 0.16, and 0.15, respectively, as per Table 2. This indicates that higher values of these enzyme features increase the probability of the classification being LD.

Figure 5 displays the feature total proteins' maximum, minimum, and interquartile range for both classes. It can be observed that the 25th, 50th, 75th, and 100th percentile values in the LD class are slightly lower than those in the No LD class, though comparable. The correlation of total proteins with LD diagnosis is -0.03, indicating minimal association. The negative correlation suggests that lower values of total proteins slightly increase the probability of an LD classification.

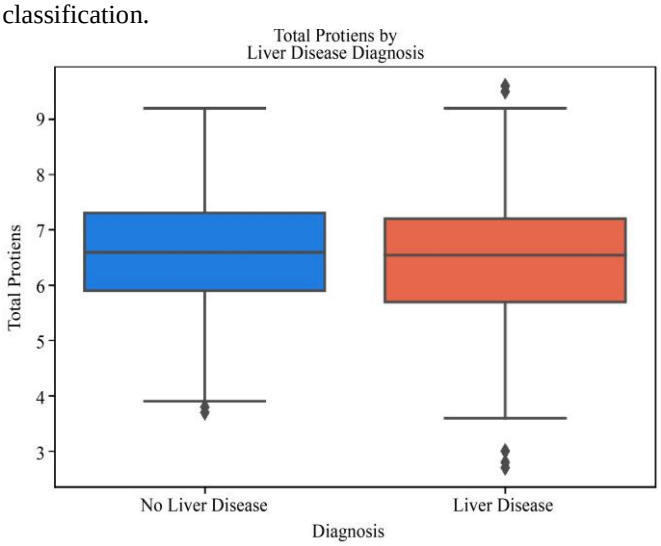

Fig. 5 Total proteins for Liver Disease classification

Figure 6 illustrates the minimum, maximum, and interquartile ranges of the albumin features, albumin and albumin to globulin ratio, for both diagnosis classes.

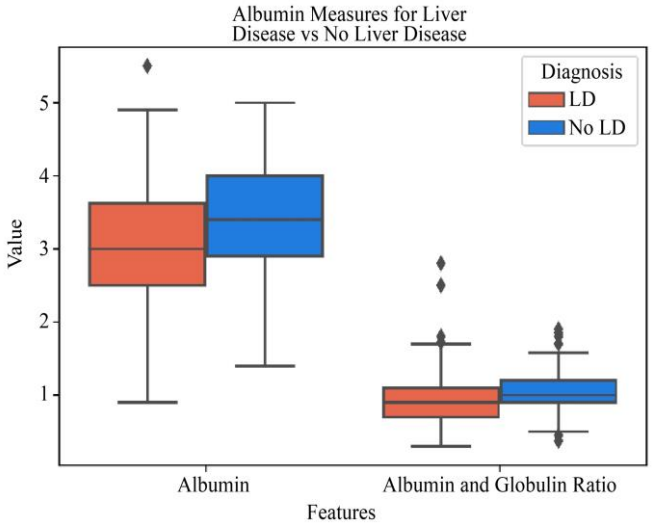

Fig. 6 Albumin for Liver Disease classification

It is observed that the 25th, 50th, 75th, and 100th percentile values of "albumin and the albumin to globulin ratio" are lower for the LD class than for the No LD class. Both "albumin and the albumin to globulin ratio" have correlation values of -0.16 (as per Table 2), indicating a moderate association with LD diagnosis. The negative correlation suggests that lower values of "albumin and the albumin to globulin ratio" increase the probability of an LD classification.

3.3. Data Balancing and Upsampling

The input dataset originally contained data from 579 patients. Figure 7 shows the original distribution of LD diagnoses in the dataset. Out of 579 entries, 415 are classified as LD and 165 as No LD.

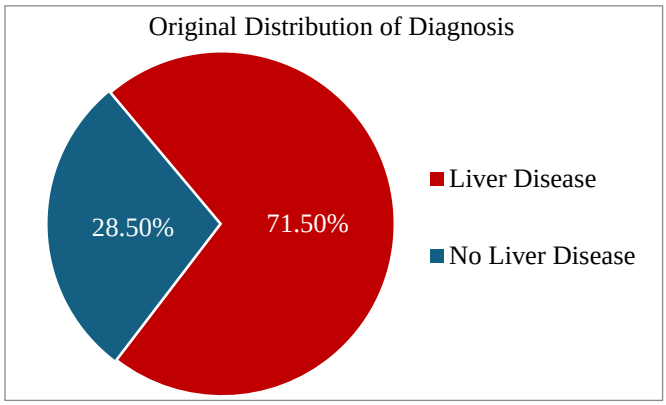

Fig. 7 Original distribution of Liver Disease dataset

This imbalance in the count of LD and No LD cases may lead the models to become biased toward a particular class. It is necessary to balance the dataset through random upsampling to address this issue. Random upsampling involves increasing the number of No LD cases to match the number of LD cases by adding random rows with a No LD diagnosis. This approach effectively balances the dataset and helps prevent bias in the models.

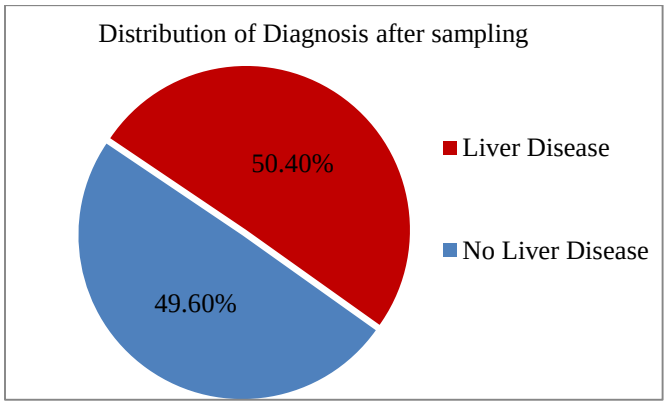

Fig. 8 Distribution of Liver Disease classes after data balancing and upsampling

Given that the dataset size is relatively small, it may not be sufficient for the models to learn complex patterns,

highlighting the need for further upsampling of the balanced dataset. For this study, we have tripled the balanced dataset through random upsampling. The distribution of LD diagnoses after data balancing and upsampling is illustrated in Figure 8.

3.4. Methods for Liver Diseases

3.4.1. Gradient Boosting (GB)

GB is a sequential ensemble of multiple decision trees acting as weak learners. Each tree minimizes the errors of the previous ones. The model updation using GB is carried out using

$$F_m(x) = F_{m-1}(x) + \eta h_m(x) \quad (1)$$

Where $F_{m-1}(x)$is the prediction from the previous trees and $h_m(x) = arg\ \min_h \sum_i \left(r_{i,m} - h(x_i)\right)^2$ is the new tree that fits the negative gradient of the loss function given by

$$r_{i,m} = \left[\frac{\partial L\left(y_i, F_{m-1}(x_i)\right)}{\partial F_{m-1}(x_i)}\right] \quad (2)$$

The new tree $h_m(x)$ is trained to minimize the squared error with respect to these residuals.

The model iteratively builds decision trees that predict whether a patient has liver disease based on their clinical features.

3.4.2. Extreme Gradient Boosting (XGB)

XGB is an efficient version of GB that prevents overfitting using regularization techniques and supports fast tree building. The objective function used in XGB is

$$L(\Theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (3)$$

Here $l(y_i, \hat{y}_i)$is the loss function with $y_i$and $\hat{y}_i$ as the actual and predicted class of Liver disease.

$\Omega(f_k) = \gamma T + \frac{1}{2}\|w\|^2$ is the regularization term for each tree f_k to prevent overfitting with T the number of leaves in the tree and w is the weight vector of the leaves.

3.4.3. Adaptive Boosting (ADB)

ADB assigns weights to misclassified instances and updates these weights iteratively to focus on difficult-to-classify cases. Each instance (x_i, y_i) is assigned an initial weight $w_i = \frac{1}{n}$ where n is the number of training samples. At m^{th} iteration, a weak classifier $h_m(x)$ is trained, and a weight $\alpha_m = \frac{1}{2} log\left(\frac{1-\varepsilon_m}{\varepsilon_m}\right)$ is assigned based on the classifier's error $\varepsilon_m = \frac{\sum_{i=1}^{n} w_i . 1\{h_m(x_i) \neq y_i\}}{\sum_{i=1}^{n} w_i}$.

Finally, the weights for misclassified samples are updated using

$$w_{i+1} = w_i\ exp(\alpha_m . 1\{h_m(x_i) \neq y_i\}) \quad (4)$$

The final model is obtained by taking a weighted sum of all weak classifiers using $H(x) = \sum_m \alpha_m h_m(x)$

3.4.4. Histogram-Based Gradient Boosting (HB)

HB works by discretizing continuous clinical features into histograms with K bins, which reduces the complexity of the feature split search. It computes the first-order derivatives (gradients) $g_i = \frac{\partial L(y_i, \hat{y}_i)}{\partial \hat{y}_i}$ and second-order derivatives (Hessians)$h_i = \frac{\partial^2 L(y_i, \hat{y}_i)}{\partial \hat{y}_i^2}$of the loss function with respect to the prediction $F(x_i)$. These derivatives are aggregated for each bin $k$rather than for each data point, significantly reducing the computational load.

$$G_k = \sum_{i \in bin_k} g_i\ and H_k = \sum_{i \in bin_k} h_i$$

After selecting the best split, the decision tree is updated by splitting the data according to the optimal thresholdθ, and the leaf values are updated. The predictions for the terminal nodes w_k based on leaf values are computed using the aggregated gradient and Hessian in each bin using

$$w_k = -\frac{G_k}{H_k + \lambda}$$

The final prediction for sample i is the sum of predictions from all trees in the ensemble

$$F(x_i) = \sum_{m=1}^{M} w_m\ .\ h_m(x_i)$$

Where $h_m(x_i)$ is the output of the m^{th} tree for input x_i, and M is the total number of trees.

3.5. Explainable AI

3.5.1. SHapley Additive exPlanations (SHAP)

SHAP is used for both global as well as local interpretation, providing insights into feature importance through Shapley values based on cooperative game theory. It quantifies the contribution of each clinical feature to the prediction of liver disease, both across the entire dataset as well as for individual cases.

For a model with function$f(x)$ at an instant$x$, the Shapley value of feature i across all possible feature subsets $S$is calculated using

$$\varphi_i(f, x) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\ (|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)] \quad (5)$$

Where:

- $F$is the set of all features.
- $f(S)$ is the model prediction from features in S
- $f(S \cup \{i\})$ is the model prediction for combined features of i and S,
- $\varphi_i(f,x)$ represents the contribution of feature i to the prediction, for instance x.

The model's prediction $f(x)$ for a given sample x can be expressed as the sum of all feature contributions (Shapley values)

$$f(x) = \varphi_0 + \sum_{i=1}^{n} \varphi_i$$

where φ_0 is the average model output (the baseline value, often the expected prediction without any features), and$\varphi_i$are the Shapley values for the individual features.

3.5.2. Local Interpretable Model-agnostic Explanations (LIME)

LIME offers local explanations by fitting a simple and interpretable model for the prediction around a specific instance. It explains predictions by focusing on feature contributions for each specific patient. It generates perturbations of the input instance and observes the changes in the model's prediction to learn which features are most important. LIME fits a local surrogate linear model $g(z')$ to approximate the complex model f around a specific instance$x$. This model is trained on a neighborhood of perturbations z' of the instance$x$. The objective is to minimize the loss function that represents the difference between the original model f and the local surrogate model g, subject to a complexity constraint on g as given by

$$arg\min_{g \in G} \sum_{z' \in Z} L\big(f(x), g(z')\big).\pi_x(z') + \Omega(g) \qquad (6)$$

Where:

$L\big(f(x), g(z')\big)$is the loss function measuring the difference between the predictions of f and g

$\pi_x(z')$ is a proximity function that assigns higher weights to instances z′ that are closer to the instance x,

$\Omega(g)$ is a regularization term that penalizes complex models g,

Z is the set of perturbed instances generated around x

Liver disease presence is predicted using the linear regression model

$$g(z') = \beta_0 + \sum_{i=1}^{n} \beta_i z'_i \qquad (7)$$

Where $z_i'$are the perturbed feature values, and β_i are the weights representing the contribution of each feature to the local approximation.

3.5.3. ELI5

ELI5 is a tool that provides model interpretations by computing feature importance scores and explaining the inner workings of machine learning models. For tree-based models, ELI5 measures feature importance by evaluating how often and effectively a feature contributes to reducing the model's loss function at different splits. The importance score for feature i is calculated using

$$\text{Importance}(x) = \sum_{splitson x_i} \frac{reductioninlossduetosplit}{totalreductioninloss} \qquad (8)$$

This reflects how influential a feature is in guiding the model toward better predictions, which is used to understand the overall importance of features in a liver disease classification model.

4. Results and Discussion

This section provides the quantitative analysis of the proposed method for liver disease classification using various metrics and the interpretation of the diagnosis using XAI.

4.1. Quantitative Metrics

The four-ensemble ML algorithms are used for liver disease classification. The following metrics are used for assessment.

- $Accuracy = \frac{TP+TN}{N}$ as the correct classification for both the presence and absence of liver disease.
- $\text{Precision} = \frac{TP}{TP+FP}$ as the predicted liver disease cases for patients that were truly positive for liver disease.
- $Re\,c\,all = \frac{TP}{TP+FN}$as the actual liver diseased cases that were correctly identified.
- $F_1 - Score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$ is the harmonic mean of precision and recall, balancing both metrics.
- Log Loss is a measure that penalizes underconfident and incorrect predictions significantly. It is expressed using the following equation for actual outcomes y_i and predicted probabilities p_i for total N cases:

  $$LogLoss = -\frac{1}{N}\left(\sum_{i=1}^{N}(y_i\, log(p_i) + (1-y_i)\, log(1-p_i))\right)$$
- Jaccard Score $JS = \frac{|y_i \cap \hat{y}_i|}{|y_i \cup \hat{y}_i|}$is the measure of similarity and intersection between actual output $y_i$and predicted output $\hat{y}_i$.
- Dice Coefficient $DC = \frac{2|y_i \cap \hat{y}_i|}{|y_i| + |\hat{y}_i|}$is the measure of overlap.
- Matthews Correlation Coefficient $MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$is the measure of difference.

Table 3 shows the prediction performance of the ensemble learning models evaluated using training and testing

accuracies. It is observed that GB produced the highest training and testing accuracies of 0.9666 and 0.9598, respectively. HB emerged as the second-best performer, having training and testing accuracies of 0.9436 and 0.9316, respectively. XGB produced training and testing accuracy values of 0.9402 and 0.9357, respectively, slightly less than those of HB. ADB produced the least training and testing accuracy values of 0.8734 and 0.8405, respectively.

Table 3. Evaluation of the ensemble learning models based on training and testing accuracies

SN	Model	Training Accuracy	Testing Accuracy
1	GB	0.9666	0.9597
2	HB	0.9436	0.9316
3	ADB	0.8734	0.8404
4	XGB	0.9401	0.9356

Table 4 compares the prediction performances of the ensemble ML models based on precision, recall, F1 Score and specificity. It can be noticed that GB outperforms every other model with a Precision of 0.9617, Recall of 0.9598, specificity of 0.9920 and F1 score of value of 0.9597. XGB stood as the second-best performer, producing a precision of 0.9402, recall of 0.9357, specificity of 0.9866 and F1 Score value of 0.9355. HB exhibited a precision of 0.9355, recall of 0.9316, specificity of 0.9786 and F1 score value of 0.9315. ADB underperformed every other model, producing precision, recall, F1 score and specificity and values of 0.8460, 0.8405, 0.8398 and 0.9035, respectively.

Table 4. Performance evaluation of the ensemble learning models based on precision, recall, F1 score and specificity

SN	Model	Precision	Recall	F1 Score	Specificity
1	GB	0.9616	0.9597	0.9597	0.9919
2	HGB	0.9354	0.9316	0.93148	0.9785
3	ADB	0.8459	0.8404	0.83984	0.9034
4	XGB	0.9402	0.9356	0.93548	0.9865

Table 5. Performance evaluation of the ensemble learning models based on JS, DC and MCC

SN	Model	JS	DC	MCC
1	GB	0.9202	0.9584	0.9214
2	HGB	0.8661	0.9282	0.86709
3	ADB	0.7090	0.8297	0.6864
4	XGB	0.8730	0.9322	0.8758

Table 5 shows the comparison of the prediction performances of these ensemble learning models based on JS, DC and MCC. It can be noticed that GB outperforms every other model in terms of all the metrics having JS, DC and MCC of 0.9202, 0.9584 and 0.9215, respectively. XGB stood as the second-best performer, producing JS, DC and MCC values of 0.8730, 0.9322 and 0.8759, respectively. HB exhibited JS, DC and MCC values of 0.8661, 0.9283 and 0.8671, respectively. ADB underperformed every other model, producing JS, DC and MCC values of 0.7090, 0.8298 and 0.6864, respectively.

Figure 9 illustrates the performance comparison of the ensemble learning models based on Log Loss. It can be observed that GB, HGB, ADB and XGB generated log loss values of 0.24, 0.28, 0.66 and 0.26, respectively. GB generated the least log loss value, indicating better performance than every other model. XGB generated a log loss of 0.2476, slightly higher than GB and less than HGB and XGB, making XGB the second-best performer in terms of log loss. HGB produced higher log loss than GB and XGB, indicating relative underperformance compared to GB and HGB. ADB produced the highest log loss and appeared the worst performer among all the classifiers considered. Figure 10 illustrates the ROC curves with AUC values for the ensemble models used in the study. It is observed that GB and XGB produced AUC values of 0.98, and HB produced an AUC value of 0.97, indicating good differentiating ability among LD and No LD cases. On the other hand, ADB produced an AUC of 0.92, indicating less differentiating ability relative to every other model.

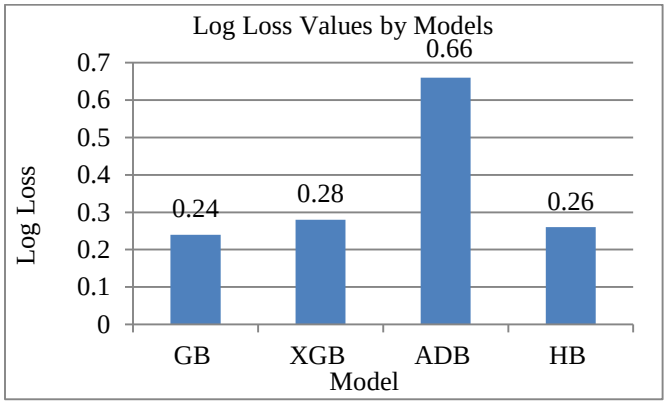

Fig. 9 Log Loss Values obtained by GB, XGB, ADB and HB for Liver Diseases classification

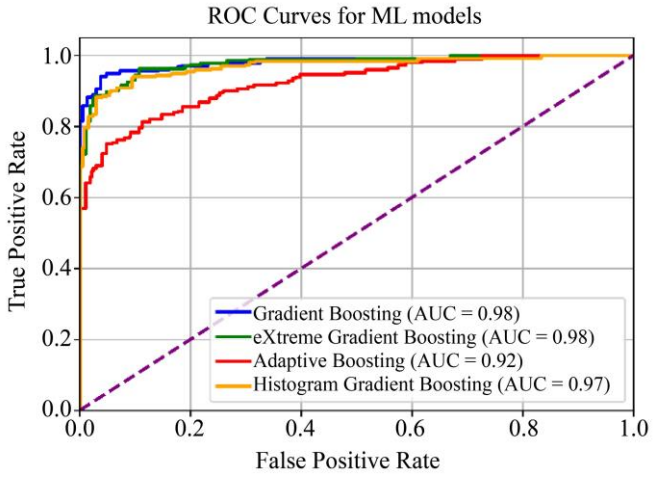

Fig. 10 ROC Curves for ML models with AUC values

Figure 11 illustrates the LIME explanation for the Gradient Boosting (GB) model applied to the 7th sample in the dataset. It can be observed that direct bilirubin, total proteins,

and gender have positive feature weights, while the remaining features have negative weights. For this instance, direct bilirubin had the highest absolute impact on the model's output, followed by alanine aminotransferase and age as the second and third most impactful features, respectively. It should be noted that LIME explanations are local and specific to individual instances; thus, these weights may not fully align with patterns in the complete dataset.

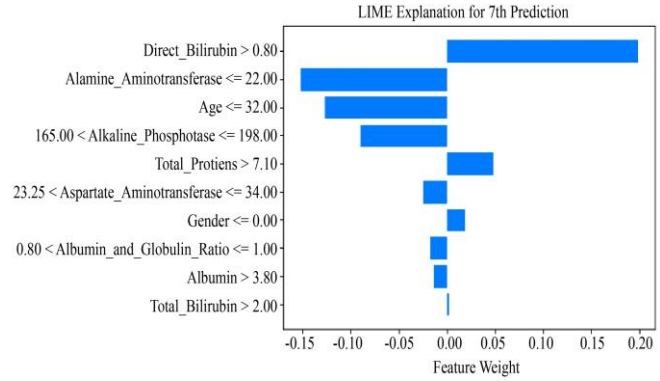

Fig. 11 LIME explanation for Liver Disease classification

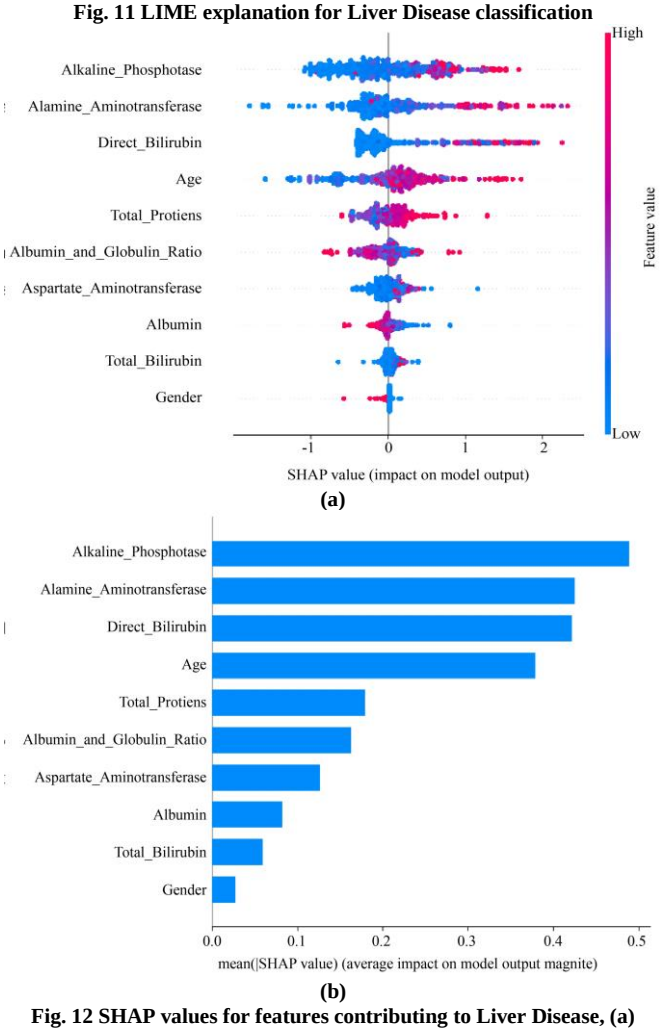

Fig. 12 SHAP values for features contributing to Liver Disease, (a) showing impact on model output, and (b) Showing mean impact on model output

Figure 12 illustrates the SHAP explanation for the GB model, displaying features and their corresponding mean absolute SHAP values. It is observed that alkaline phosphatase has the highest mean absolute SHAP value among all the features, indicating it has the greatest impact on the model's predictions. Direct bilirubin follows with the second-highest mean absolute SHAP value, signifying it is the second most significant feature in the prediction process.

Similarly, "alanine aminotransferase and age" rank third and fourth, respectively, in terms of mean absolute SHAP values. Thus, "alkaline phosphatase, direct bilirubin, alanine aminotransferase, and age" are the four most impactful features. "Aspartate aminotransferase, total proteins, albumin, albumin to globulin ratio, and total bilirubin" demonstrate moderate impacts on the model's output. It is important to note that gender has an almost negligible mean absolute SHAP value, indicating it has minimal impact on the model's predictions.

Table 6 provides the ELI5 explanation of the Gradient Boosting (GB) model, listing features and their corresponding weights. ELI5 weights are expressed as X ± Y, where X represents the mean (or estimated) weight, and Y denotes the standard deviation, a measure of variability. A higher estimated weight indicates a greater impact on the model's output. Direct bilirubin has the highest estimated weight of 0.2048, indicating the strongest influence on the model's predictions. Enzymes alkaline phosphatase, alanine aminotransferase, and age have estimated weights of 0.2045, 0.1621, and 0.1119, respectively, signifying a high impact on the output. In contrast, gender has an estimated weight of only 0.0164, indicating minimal to no impact on the model's output.

Table 6. Feature weights for Liver Disease using ELI5

SN	Feature	Weight
1	Alkaline Phosphotase	0.2161 ± 0.2015
2	Direct Bilirubin	0.2043 ± 0.2089
3	Alamine Aminotransferase	0.1212 ± 0.1751
4	Age	0.1205 ± 0.2197
5	Aspartate Aminotransferase	0.0918 ± 0.2847
6	Albumin	0.0815 ± 0.1299
7	Total Protiens	0.0683 ± 0.1062
8	Albumin and Globulin Ratio	0.0495 ± 0.1215
9	Total Bilirubin	0.0401 ± 0.1223
10	Gender	0.65 .0601

5. Conclusion

This paper effectively proposes an ensemble-based explainable ML framework for LD classification, combining robust quantitative performance with explainability, which enhances its suitability for clinical applications. The data balancing and upsampling approach also prevented model bias towards any class, resolving challenges related to dataset size limitations. After training on the balanced and upsampled dataset, Gradient Boosting (GB) emerged as the best

performer across all metrics, achieving training and testing accuracies of 0.9666 and 0.9597, respectively. GB also demonstrated strong results in Precision, Recall, F1 Score and specificity, with values of 0.9616, 0.9597, 0.9597 and 0.9919, respectively. Furthermore, the model achieved a Jaccard Score of 0.9202, a Dice Coefficient of 0.9584, and a Matthews Correlation Coefficient of 0.9214. With an impressive ROC AUC value of 0.98, GB showed excellent differentiating capability between LD and No LD classes. Additionally, GB achieved the lowest log loss among all ensemble models evaluated in this study, with a value of 0.24, underscoring its reliability in LD classification. Through explanations provided by SHAP, LIME, and ELI5, it was found that Direct Bilirubin, Alkaline Phosphatase, Alanine Aminotransferase, and Age are the most impactful features on the model's output, indicating these as highly influential factors in classification.

References

[1] Peter Byass, "The Global Burden of Liver Disease: A Challenge for Methods and for Public Health," *BMC Medicine*, vol. 12, pp. 1-3, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[2] Liver Disease, Cleveland Clinic, 2024. [Online]. Avaliable: https://my.clevelandclinic.org/health/diseases/17179-liver-disease

[3] Neda Afreen et al., "A Novel Machine Learning Approach using Boosting Algorithm for Liver Disease Classification," *5th International Conference on Information Systems and Computer Networks*, Mathura, India, pp. 1-5, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[4] G. Shobana, and K. Umamaheswari, "Prediction of Liver Disease Using Gradient Boost Machine Learning Techniques with Feature Scaling," *5th International Conference on Computing Methodologies and Communication*, Erode, India, pp. 1223-1229, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[5] Vinayak Singh, Mahendra Kumar Gourisaria, and Himansu Das, "Performance Analysis of Machine Learning Algorithms for Prediction of Liver Disease," *IEEE 4th International Conference on Computing, Power and Communication Technologies*, Kuala Lumpur, Malaysia, pp. 1-7, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[6] Krittika Dutta, Satish Chandra, and Mahendra Kumar Gourisaria, "Early-Stage Detection of Liver Disease through Machine Learning Algorithms," *Advances in Data and Information Sciences*, pp. 155-166, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[7] Anton Sokoliuk et al., "Machine Learning Algorithms for Binary Classification of Liver Disease," *IEEE International Conference on Problems of Infocommunications, Science and Technology*, Kharkiv, Ukraine, pp. 417-421, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[8] Mounita Ghosh et al., "A Comparative Analysis of Machine Learning Algorithms to Predict Liver Disease," *Intelligent Automation & Soft Computing*, vol. 30, no. 3, pp. 917-928, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[9] Fazle Rabbi et al., "Prediction of Liver Disorders Using Machine Learning Algorithms: A Comparative Study," *2nd International Conference on Advanced Information and Communication Technology*, Dhaka, Bangladesh, pp. 111-116, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[10] Ketan Gupta et al., "Liver Disease Prediction Using Machine Learning Classification Techniques," *IEEE 11th International Conference on Communication Systems and Network Technologies*, Indore, India, pp. 221-226, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[11] Greeshma Arya et al., "Explainable AI for Enhanced Interpretation of Liver Cirrhosis Biomarkers," *IEEE Access*, vol. 11, pp. 123729-123741, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[12] A. Nilofer, and S. Sasikala, "A Comparative Study of Machine Learning Algorithms Using Explainable Artificial Intelligence System for Predicting Liver Disease," *Computing Open*, vol. 1, pp. 1-18, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[13] Abrar Alotaibi et al., "Explainable Ensemble-Based Machine Learning Models for Detecting the Presence of Cirrhosis in Hepatitis C Patients," *Computation*, vol. 11, no. 6, pp. 1-17, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[14] Xieyi Pei et al., "Machine Learning Algorithms for Predicting Fatty Liver Disease," *Annals of Nutrition and Metabolism*, vol. 77, no. 1, pp. 38-45, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[15] Basile Njei et al., "An Explainable Machine Learning Model for Prediction of High-Risk Nonalcoholic Steatohepatitis," *Scientific Reports*, vol. 14, pp. 1-9, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[16] Ebenezer Agbozo, and Daniel Musafiri Balu, "Liver Disease Classification-An XAI Approach to Biomedical AI," *Informatica*, vol. 48, no. 1, pp. 1-12, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[17] Cristina Baciu et al., "Artificial Intelligence Applied to Omics Data in Liver Diseases: Enhancing Clinical Predictions," *Frontiers in Artificial Intelligence*, vol. 5, pp. 1-10, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[18] R. Pandi Selvam et al., "Explainable Artificial Intelligence with Metaheuristic Feature Selection Technique for Biomedical Data Classification," *Biomedical Data Analysis and Processing Using Explainable (XAI) and Responsive Artificial Intelligence (RAI)*, pp. 43-57, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[19] Mamatha Bhat et al., "Artificial Intelligence, Machine Learning, and Deep Learning in Liver Transplantation," *Journal of Hepatology*, vol. 78, no. 6, pp. 1216-1233, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[20] Ramisetty Kavya et al., "Machine Learning and XAI Approaches for Allergy Diagnosis," *Biomedical Signal Processing and Control*, vol. 69, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[21] Seyedeh Neelufar Payrovnaziri et al., "Explainable Artificial Intelligence Models Using Real-World Electronic Health Record Data: A Systematic Scoping Review," *Journal of the American Medical Informatics Association*, vol. 27, no. 7, pp. 1173-1185, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[22] Ruey-Kai Sheu, and Mayuresh Sunil Pardeshi, "A Survey on Medical Explainable AI (XAI): Recent Progress, Explainability Approach, Human Interaction and Scoring System," *Sensors*, vol. 22, no. 20, pp. 1-42, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[23] Bendi Ramana, and N. Venkateswarlu, "ILPD (Indian Liver Patient Dataset)," *UC Irvine Machine Learning Repository*, 2022. [CrossRef] [Publisher Link]