Enhanced Liver Disease Prediction Using Machine Learning Algorithms: Implementation, Evaluation, and Improvements

M.Reagan

Department of MCA, School of engineering,
Chanakya university, Bengaluru-562165

Gore Shriraj Laxman

Department of MCA, School of engineering,
Chanakya university, Bengaluru-562165

Abstract

“This study enhances a previous machine learning approach for liver disease prediction using the Indian Liver Patient Dataset (ILPD). The original work applied Logistic Regression, K-Nearest Neighbours, and SVM, achieving test accuracies around 73–74%, but was limited by a narrow model scope and minimal evaluation. To enhance this, we re-implemented the base models and added more classifiers such as Decision Tree, Random Forest, Naive Bayes, Gradient Boosting and XGBoost. a With 70-30 train-test splitting, we measured both accuracy and AUC-ROC and examined feature importance from ensemble models. Using a 70-30 train-test split, we evaluated both accuracy and AUC-ROC and analysed feature importance from ensemble models. Results showed Logistic Regression and SVM achieved the highest test accuracy (73.71%), while ensemble methods offered competitive performance and better interpretability. Overall, our expanded approach provides a more robust and clinically insightful evaluation of liver disease prediction models.”

Keywords: Liver Disease, Machine Learning, ROC Curve, Logistic Regression, KNN, Support Vector Machine

I. INTRODUCTION

The liver is one of the most important organs in the human body. It helps with digestion, removes toxins from the blood, and plays a key role in many other vital functions. If the liver gets damaged and the problem is not detected early, it can lead to serious health issues or even be life-threatening.

In recent years, machine learning (ML) has been used to help doctors detect diseases like liver problems by analysing medical data. However, many of these studies have some weaknesses. They often do not clean or prepare the data well, use only a few basic models, or do not measure how accurate or useful the results really are.

This research builds on a previous study that used three basic models: Logistic Regression, K-Nearest Neighbour (KNN), and Support Vector Machine (SVM). While these models gave some results, they were not very accurate or reliable. In our work, we fix the problems from the earlier study. We use better methods to prepare the data, test more advanced models, and compare their results in a detailed way. This helps us find which models work best for predicting liver disease.

III. DATASET AND PREPROCESSING

A. Dataset Description

The dataset used for this study is the Indian Liver Patient Dataset (ILPD), which was obtained from the UCI Machine Learning Repository. The dataset consists of 583 individual medical records of patients from Andhra Pradesh, India. Each record is comprised of 10 biological attributes and 1 target variable that indicates whether the patient is diagnosed with a liver disease.

Attribute	Description
Age	Patient age in years
Gender	Male or Female
Total_Bilirubin	Total bilirubin level (mg/dL)
Direct_Bilirubin	Direct bilirubin level (mg/dL)
Alkaline_Phosphotase	ALP enzyme level (U/L)
Alamine_Aminotransferase	ALT enzyme level (U/L)
Aspartate_Aminotransferase	AST enzyme level (U/L)
Total_Proteins	Total protein in blood (g/dL)
Albumin	Albumin level (g/dL)
Albumin_and_Globulin_Ratio	Ratio of albumin to globulin
Liver_Disease (Target)	1 = Disease, 2= No Disease

The dataset is inherently imbalanced, with a greater number of records labeled as having liver disease (1) compared to the healthy class (2). Therefore, accuracy alone is not a sufficient evaluation metric — metrics such as sensitivity, specificity, and ROC-AUC were considered to assess the models more effectively.

B. Data Cleaning and Feature Engineering

To ensure the quality of data before feeding it into machine learning models, a comprehensive data preprocessing pipeline was developed, as follows:

1) Target Label Encoding

The original dataset labelled the target variable as 1 for patients with liver disease and 2 for those without. To align with standard binary classification practices in machine learning where 1 typically represents the positive class and 0 the negative these labels were remapped accordingly: 1 indicating the presence of liver disease and 0 indicating its absence.

2) Categorical Encoding

The Gender attribute, which was originally categorical with values "Male" and "Female," was converted into a numeric binary format using label encoding—assigning 1 to Male and 0 to Female. This transformation ensures compatibility with machine learning algorithms like Logistic Regression and SVM, which require numerical input rather than text.

3) Handling Missing Values

The dataset was analysed for null and missing values. A few entries in Albumin_and_Globulin_Ratio were missing and were imputed using the mean value of the respective column, as it retains the central tendency without skewing distribution.

4) Feature Scaling

As many features (e.g., bilirubin, enzyme levels) had skewed distributions, they were normalized using StandardScaler to ensure zero mean and unit variance. This scaling step is crucial for distance-based models (e.g., KNN) and gradient-based optimizers in SVM or Logistic Regression.

5) Correlation Analysis

To assess potential multicollinearity among the input features, a correlation matrix was generated. As anticipated, there was a strong positive correlation between Total Bilirubin and Direct Bilirubin, since the latter is a component of the former. Additionally, a significant linear relationship was observed between the liver enzymes AST (Aspartate Aminotransferase) and ALT (Alamine Aminotransferase), which often rise together in cases of liver dysfunction.

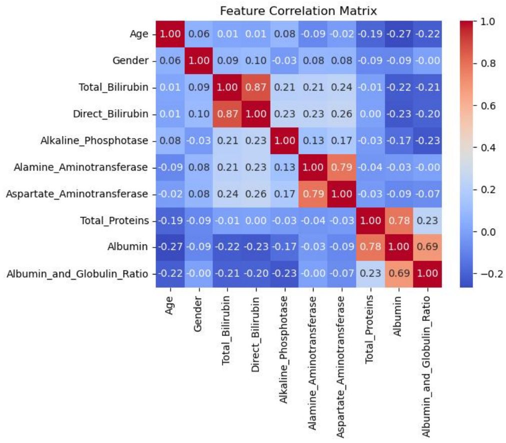


Figure 1: Heatmap illustrating the Pearson correlation between numerical attributes.

The heatmap helps in identifying redundant features and understanding inter-variable relationships, which is useful for feature selection and model interpretability.

C. Exploratory Data Visualization

To further understand class distribution and feature interaction, a pair plot was generated using the Seaborn library.

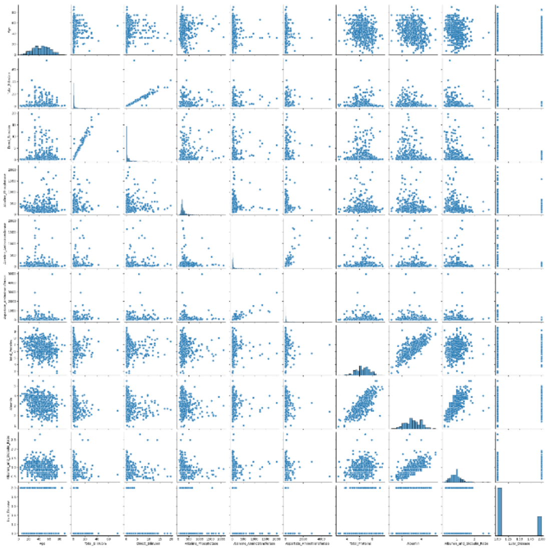


Figure 2.1.

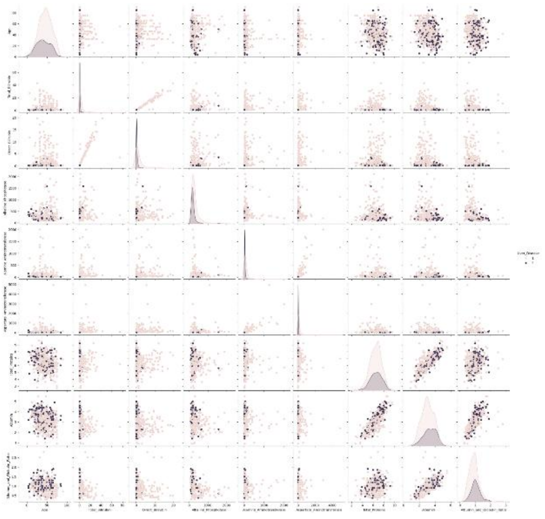


Figure2.2 Pair plot showing feature relationships across liver disease and non-disease cases

The pair plot shows that features like *Total Bilirubin*, *Direct Bilirubin*, and liver enzymes are generally higher in patients with liver disease, indicating some class separation. Strong correlations between *Total* and *Direct Bilirubin*, and among liver enzymes, suggest feature redundancy. The plot also confirms class imbalance, with more cases of liver disease. Several features are right-skewed, highlighting the need for normalization. Lastly, the lack of clear linear boundaries supports the use of ensemble or kernel-based models over simple linear classifiers.

V. METHODOLOGY

In this study, we adopted a comprehensive machine learning approach for the prediction of liver disease, improving upon an earlier study by integrating more robust preprocessing, additional models, and modern evaluation strategies. The methodology is organized into two main components: the algorithms used and the processing pipeline.

A. Classification Models

1) Existing Models

The following models were originally implemented in the prior research and were re-evaluated in our enhanced study:

- **Logistic Regression**: A linear model used for binary classification that estimates the probability of class membership using the sigmoid function..
- **K-Nearest Neighbours (KNN)**:A non-parametric algorithm that classifies data points based on the majority class of the k nearest data points.
- **Support Vector Machine (SVM)**:A powerful classifier that aims to find the optimal separating hyperplane with the maximum margin between classes.

These models served as the baseline for comparative evaluation. While effective, they were limited in handling feature interactions and imbalanced datasets, and lacked scalability for more complex patterns.

2) Models Implemented

To address the limitations of the baseline models, we incorporated more sophisticated algorithms into the study:

- **Random Forest**: An ensemble of decision trees trained on random subsets of data and features. It reduces overfitting, handles missing values, and provides feature importance.
- **XG Boost (Extreme Gradient Boosting)**:A scalable and regularized gradient boosting framework that sequentially improves the model by correcting the previous errors.
- **Gradient Boosting:** A boosting technique that builds models sequentially, where each new model attempts to correct the errors made by the previous ones. Unlike XGBoost, it does not include advanced regularization by default but still performs well in capturing complex patterns.
- **Naive Bayes**:A probabilistic model based on Bayes' theorem assuming independence between features.

These models were new additions to the framework and were selected based on their demonstrated success in biomedical classification problems, especially where feature interactions and non-linear patterns are important.

B. Processing Pipeline

The complete machine learning pipeline used in this study followed these steps:

1. **Label Encoding and Missing Value Imputation**
   The Gender feature, originally in text form (Male/Female), was converted to numerical format using label encoding. Missing values, particularly in the Albumin and Globulin Ratio column, were filled using the mean of the respective column.
2. **Train-Test Splitting**
   The dataset was split into training and testing sets using a 70:30 ratio. Stratified sampling was applied to maintain a balanced distribution of the target classes in both sets
3. **Feature Normalization**
   All numerical features were standardized using the StandardScaler technique. This step ensured that features with larger scales didn't dominate distance-based or gradient-based models like KNN or SVM.
4. **Model Evaluation**
   Each model was evaluated using five key metrics to assess performance comprehensively:
   - **Accuracy**: overall correctness of predictions
   - **Confusion Matrix**: detailed breakdown of true/false predictions
   - **Sensitivity (Recall)**: how well the model detects actual disease cases
   - **Specificity**: how well the model identifies healthy individuals
   - **ROC-AUC**: a threshold-independent measure of overall classification quality

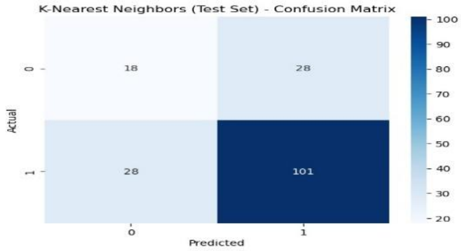
V. RESULTS AND DISCUSSION

The table summarizes the performance of all evaluated machine learning models based on four key metrics: Accuracy, ROC-AUC, Sensitivity, and Specificity**.**

Table 1: Performance Metrics Table

Model	Accuracy	ROC-AUC	Sensitivity	Specificity
Logistic Regression	73.71%	0.78	0.91	0.24
KNN	68.00%	0.76	0.78	0.39
SVM	73.71%	0.75	1.00	0.00
Decision Tree	64.00%	0.70	0.76	0.30
Random Forest	72.00%	0.88	0.85	0.35
Naive Bayes	57.71%	0.77	0.43	0.98
XGBoost	72.57%	0.91	0.84	0.41
Gradient Boosting	70.29%	0.88	0.84	0.30

All the classification algorithms were tested. Their confusion matrices and derived metrics are explained below:



:

Figure 3: Confusion matrix of K-Nearest Neighbour. From the confusion matrix of the K-Nearest Neighbour model, accuracy was calculated as 68.00%. The sensitivity value was 0.78, and the specificity value was 0.39.

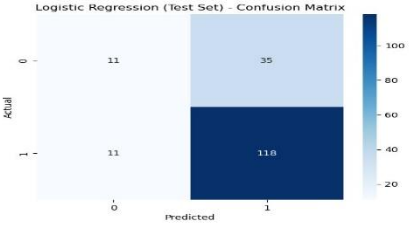


Figure 4: Confusion matrix of Logistic Regression. From the confusion matrix of the Logistic Regression model, accuracy was calculated as 73.71%. The sensitivity value was 0.91, and the specificity value was 0.24.

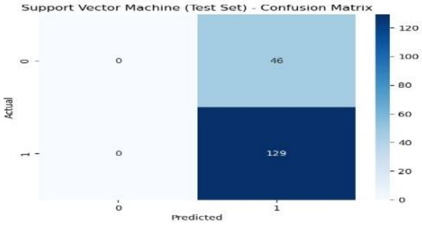


Figure 5: Confusion matrix of Support Vector Machine. From the confusion matrix of the Support Vector Machine model, accuracy was calculated as 73.71%. The sensitivity value was 1.00, and the specificity value was 0.00.

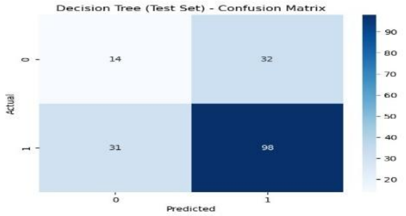


Figure 6: Confusion matrix of Decision Tree. From the confusion matrix of the Decision Tree model, accuracy was calculated as 64.00%. The sensitivity value was 0.76, and the specificity value was 0.30.

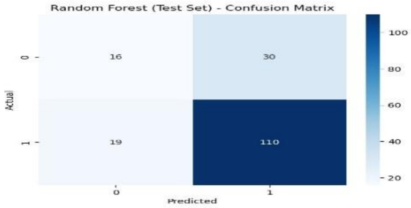


Figure 7: Confusion matrix of Random Forest. From the confusion matrix of the Random Forest model, accuracy was calculated as 72.00%. The sensitivity value was 0.85, and the specificity value was 0.35.

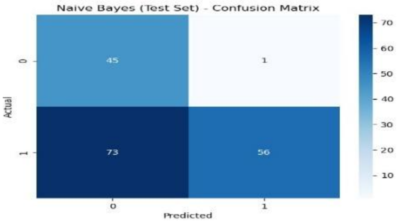


Figure 8: Confusion matrix of Naive Bayes. From the confusion matrix of the Naive Bayes model, accuracy was calculated as 57.71%. The sensitivity value was 0.43, and the specificity value was 0.98.

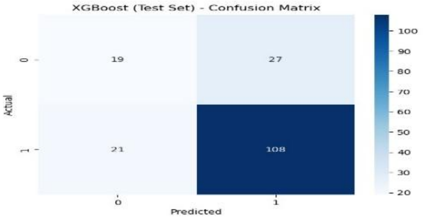


Figure 9: Confusion matrix of XGBoost. From the confusion matrix of the XGBoost model, accuracy was calculated as 72.57%. The sensitivity value was 0.84, and the specificity value was 0.41.

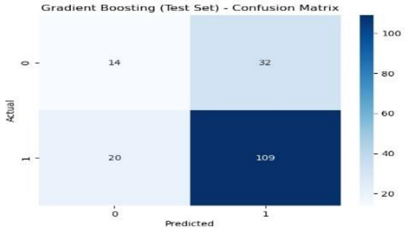


Figure 10: Confusion matrix of Gradient Boosting. From the confusion matrix of the Gradient Boosting model, accuracy was calculated as 70.29%. The sensitivity value was 0.84, and the specificity value was 0.30.

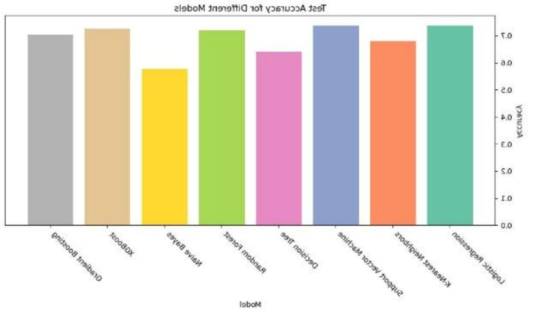


Figure 11: Accuracy comparison bar plot across all models.

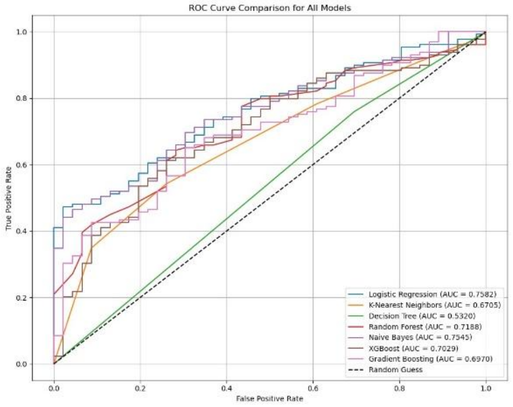


Figure 12: ROC Curve comparison across all models.

In addition to sensitivity and specificity, the ROC (Receiver Operating Characteristic) curve and its associated Area Under the Curve (AUC) offer a robust evaluation of model performance. The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1 - specificity) across different classification thresholds. AUC quantifies this performance; a score close to 1.0 indicates excellent class separation, whereas 0.5 suggests random guessing. Among the models evaluated, XGBoost achieved the highest AUC score of 0.91, reflecting its strong discriminatory ability.

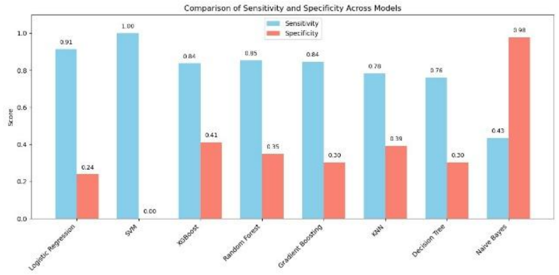


Figure 13: Comparison of sensitivity and specificity of different models.

From this table, it is clear that Logistic Regression and SVM deliver high sensitivity but at the cost of poor specificity, making them prone to false positives. SVM, in particular, labels all samples as positive, resulting in perfect sensitivity (1.00) but zero specificity (0.00). Naive Bayes stands out with an unusually high specificity (0.9783), yet it significantly underperforms on sensitivity (0.4341), failing to identify most disease cases. XG Boost offers the best trade-off with relatively high accuracy (0.7257), strong sensitivity (0.8372), and the highest specificity among the well-balanced models (0.4130). This balance reinforces its practical utility in clinical applications, where minimizing both false negatives and false positives is critical.

CONCLUSION

This study presents an improved framework for liver disease prediction by addressing key limitations in a previous baseline model. Enhancements included robust data preprocessing—such as handling missing values, encoding, and normalization—and the integration of advanced models like Random Forest and XGBoost.

Visual diagnostics, including heatmaps and pair plots, provided deeper insights into feature interactions. Evaluation using accuracy, confusion matrices, and ROC-AUC confirmed that XGBoost offered the best balance of sensitivity and specificity, making it highly suitable for clinical applications.

Ensemble methods not only improved predictive performance but also reduced overfitting, ensuring greater model reliability. Future work could incorporate explainable AI (e.g., SHAP, LIME) for better transparency and explore deep learning or hybrid models, along with more diverse datasets, to enhance generalizability.

REFERENCE

[1] UCI Machine Learning Repository: ILPD Dataset

[2] Scikit-learn Documentation. https://scikit-learn.org

[3] T. K. Thirunavukkarasu et al., "Prediction of Liver Disease using Classification Algorithms," ICCCA, 2018.

[4] XGBoostDocumentationhttps://xgboost.readthedocs.io

[5] A. Charleonnan,T. Fufaung, T. Niyomwong, W Chokchueypattanakit, S. Suwannawach, N. Ninchawee "Predictive Analytics for Chronic Kidney Disease Using Machine Learning Techniques " *MITiCON2016.*

[6] Anatomy and function of the liver. [Online]. Available: https://www.medicinenet.com/liver_anatomy_and_function/article.ht m

[7] Abhishek Chowdhury, Thirunavukkarasu K, Sidhyant Tejas(2017), Predicting whether song will be hit using Logistic Regression.

Volume 6 Issue 9 September 2017.

[8] K-Nearest Neighbours. [Online]Available:

https://www.saedsayad.com/k_nearest_neighbors.htm

[9] P.Mazaheri, A. Narouziand A. Karimi (2015), Using Algorithms to Predict Liver Disease Classification,Electronics Information and Planning. 3:255-259.

[10] H. Jin , S. Kim and J.Kim (2014), Decision Factors on Effective Liver Patient Data Prediction,International Journal of Bio-Science and Bio-Technology

[11] Available: http://www.ics.uci.edu/~mlearn/databases/

[12] Comparative Study of Artificial Neural Network based

Classification of Liver Patient, Journal of Information Engineering and Application. 3(4).

[13] David Diez, Christopher Barr and Mine Cetinkaya-Rundel. Open

Intro Statistics. 3^{rd} Edition. OpenIntro.org

[14] Reetu and N.Kumar (2015), Medical Diagnosis for Liver Cancer Using Classification Techniques, International Journal of Recent Scientific. Volume 6. Issue, 6, pp 4809-4813.

[15] Tina R. Patil, Mrs. S. S. Sherekar. Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification.

International Journal Of Computer Science And Applications Vol. 6,

No.2, Apr 2013