A Machine Learning-Based Approach for Predicting Liver Disease and Identifying Significant Risk Factors

Submitted By

Sk. Rasel Islam
(201-35-604)
Department of Software Engineering

Supervised By

Mr. Md Rajib Mia
Lecturer
Department of Software Engineering

A thesis submitted with substantial completion of the prerequisites for the Bachelor of Science in Software Engineering degree

Fall 2023

©All right reserved by Daffodil International University

APPROVAL

This thesis titled on "**A machine learning-based approach for predicting liver disease and identifying significant risk factors.**", submitted by **Sk. Rasel Islam (ID: 201-35-604)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

10/7/24

Chairman

Dr. Engr. AKM Masum
Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Marzia
10/7/24

Internal Examiner 1

Dr. Marzia Ahmed
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

Md. Rajib Mia
Lecturer
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

External Examiner

Dr. Md. Manowarul Islam
Associate Professor
Dept. of Computer Science and Engineering
Jagannath University

© Daffodil International University

DECLARATION

I announce that I am rendering this study document under Mr. Md Rajib Mia, Lecturer, Department of Software Engineering, Daffodil International University. I therefore, state that this work or any portion of it was not proposed here therefore for Bachelor's degree or any graduation.

Supervised By

Mr. Md Rajib Mia
Lecturer
Department of Software Engineering
Daffodil International University

Submitted by

Sk. Rasel Islam
ID: 201-35-604
Department of Software Engineering
Daffodil International University

© Daffodil International University

ACKNOWLEDGEMENT

The work I have completed is completely driven by my passion for information and desire to expand my studies. The study used a cutting-edge machine learning technique to predict liver disease and discover significant risk variables. Firstly, I show my thankfulness to the Almighty, Allah, who has supplied me with the wisdom to gain knowledge and participate in morally pure deeds. Furthermore, I am extremely grateful to my parents for their essential contribution to establishing my current circumstances. I would like to extend my sincere gratitude to **Prof. Dr. Imran Mahmud**, the esteemed Head of the Department of Software Engineering. To all the excellent educators who led me along my educational path. I am glad to have them as my professors and mentors. Once again, I offer my profound appreciation to my research supervisor, **Mr. Md. Rajib Mia**, for his enormous advice and consistent support during the course of this project. I strongly value his experience and wide knowledge, and I am truly appreciative of his generous desire to share them with me.

Finally, I would like to extend my appreciation to my peers and members of DIU for their exceptional engagement and assistance, which significantly contributed to my successful achievement of this goal.

© Daffodil International University

ABSTRACT

Liver illnesses constitute a serious worldwide health issue that are impacted by various risk factors, including genetic, environmental, and lifestyle components. Early identification is crucial to avoid serious health repercussions. This article proposes a machine learning-based approach to detect important risk variables and anticipate liver disease using SHAP (SHapley Additive exPlanations) values. Using a vast dataset, we utilize a number of machine learning algorithms, such as Random Forest, K-Nearest Neighbor, Support Vector Machines, Logistic Regression, Navie Bayes, Decision Tree, and XGBoost, to boost prediction accuracy. The procedure includes tight data preparation, feature selection, and model evaluation applying performance criteria including precision, recall, F1-score and accuracy. With an average ROC-AUC rate of 99.9%, the XGBoost model fared the best, closely followed by the Random Forest model with a ROC-AUC rate of 99.3%. SHAP values are used to interpret each characteristic's contribution and provide information about the important risk variables. Our findings suggest that incorporating SHAP data considerably boosts the model's interpretability and performance. The objective of this study is to give clinicians a reliable tool for early identification of individuals who are at risk, so that timely and personalized drugs may be administered to lessen the symptoms of liver problems.

© Daffodil International University

Table Of Contents

ACKNOWLEDGEMENT iv

ABSTRACT v

Chapter 1 *1*

Introduction 1

Chapter 2 *3*

Literature Review 3

Chapter 3 *5*

Methodology 5

3.1 Data Preparation 5

3.1.1 Data Collection 5

3.1.2 Data Splitting 6

3.1.3 Data Preprocessing 6

3.1.4 Data Balancing 7

3.2 Application of Machine Learning 7

3.2.1 Train Test Method 7

3.2.1.1 Decision Tree 7

3.2.1.2 Support Vector Machine 8

3.2.1.3 K-Nearest Neighbors (KNN) 8

3.2.1.4 Naive Bayes (NB) 8

3.2.1.5 Random Forest 9

3.2.1.6 LightGBM (LGB) 9

3.2.1.7 XGBoost (XGB) 10

3.2.1.8 Logistic Regression 10

3.2.2 Cross-Validation Method 11

Chapter 4 *12*

Result and Discussion 12

4.1 Segmentation of Liver and Finding Disease 12

4.1 Performance Evaluation 14

4.3 Risk Factor Evaluation and Model Interpretation Using SHAP Value to Predict Liver Disease 17

Chapter 5 *19*

Conclusion 19

REFERENCES *19*

© Daffodil International University

Chapter 1

Introduction

Liver diseases cover a wide range of conditions, including hepatitis, fatty liver disease, cirrhosis, and liver cancer, which collectively contribute to significant morbidity and mortality worldwide. The liver's essential functions, such as detoxification, metabolism, and synthesis of important proteins, make it a critical organ whose impairment can lead to severe health consequences. Early detection and accurate prediction of liver disease are important for effective management and treatment, possibly preventing the progression to more severe stages. [1]

Traditional diagnostic approaches for liver illnesses frequently rely on intrusive treatments, such as biopsies, or indirect measurements, such blood biomarkers and imaging investigations. While these methods are valuable, they have limitations in terms of accessibility, cost, and early detection capabilities. The advent of machine learning offers a viable alternative by exploiting big datasets to identify relationships and trends that may not be visible through traditional methods.[2]

This study expands upon the foundation provided by prior research, particularly the basis publication "Liver Disease Prediction and Classification using Machine Learning," which revealed the promise of machine learning approaches in predicting liver illnesses. In this research, we intend to significantly strengthen the prediction capacities and interpretability of machine learning models by adding SHapley Additive exPlanations (SHAP) values to discover major risk variables related with liver illnesses. SHAP values provide a consistent measure of feature relevance, giving insights into the impact of each risk factor in the model's predictions.[3]

Our approach involves a comprehensive analysis of different machine learning algorithms, including Logistic Regression, K-Nearest Neighbor, Support Vector Machines, and Random Forest, applied to a robust dataset.[4] The process includes meticulous data preprocessing, feature selection, and model evaluation using standard performance measures such as accuracy, precision, recall, and F1-score. By integrating SHAP values, we aim to improve the transparency and reliability of our models, making them more useful for clinical applications.[5]

© Daffodil International University

The core goals of this project are twofold: to construct a machine learning-based tool that reliably predicts liver disease and find those most relevant risk factors leading to these predictions. This dual focus not only improves the predictive power of the models but also provides actionable insights for healthcare professionals, allowing early intervention and personalized treatment strategies. Through this study, we aspire to add to the ongoing efforts in combating liver diseases and improving patient outcomes.[6]

© Daffodil International University

Chapter 2

Literature Review

The use of machine learning techniques to predict liver disease has generated a lot of attention in the rapidly developing sector of healthcare. To give readers a thorough grasp of the developments and approaches used in this field, this overview of the literature contains a synopsis of the most important results from recent investigations.

A. Using Artificial Intelligence to Forecast Liver Disease:

A variety of machine learning algorithms have demonstrated encouraging gains in diagnostic precision in the prediction of liver diseases. Convolutional neural networks (CNNs) can identify non-alcoholic fatty liver disease (NAFLD) by examining images from liver biopsies, according to Arjmand et al. (2019). They used their CNN architecture-based method to diagnose hepatic steatosis with 92.4% accuracy. [7]

Auxilia (2018) examined the efficacy of many machine learning methods in forecasting liver illness in Indian populations. Artificial neural networks (ANN), random forests, decision trees, and Naive Bayes were all used in the study. With an accuracy of 85.7%, the random forest method performed better than the others. [8]

B. Thorough Evaluations and Methodological Variations:

The study conducted by Spann et al. (2020) concentrated on the use of machine learning in liver illness and transplantation. This study focused on feature selection and preprocessing techniques to improve performance. These techniques included a variety of models, including ensemble approaches, logistic regression, and support vector machines (SVM). The study showed that accuracy levels of up to 88.9% can be achieved with ensemble techniques, and random forests in particular. [9]

Sontakke et al. (2017) discovered that while logistic regression and random forests outperformed in terms of recall, with accuracies of 85.3% and 87.6%, respectively, Naive Bayes had the best precision, at 90.1%. [10] The study examined various machine learning methods for liver disease diagnosis.

© Daffodil International University

C. Methods for Feature Selection and Data Mining:

Using a dataset on liver diseases, Baitharu and Pani (2016) studied data mining strategies for healthcare decision assistance. They investigated feature selection techniques including particle swarm optimization (PSO) to improve prediction accuracy. Using PSO-enhanced decision trees, they achieved 83.5% accuracy.[11]

Preprocessing techniques like normalization and oversampling to balance datasets were highlighted by Priya et al. (2018). With an accuracy of 86.2%, the random forest algorithm fared the best in their trial when it came to predicting liver disorders. [12]

D. Innovative Methods and Upcoming Paths:

Novel approaches to enhance the prediction of liver illness have been examined in recent research. A hybrid feature extraction paradigm was presented by Acharya et al. (2012) for the use of ultrasound images in the classification of fatty liver disease. With a 91.8% accuracy rate, our hybrid model combined morphological and textural data. [13]

In their 2018 study, Nahar and Ara examined decision tree methods for predicting hepatic illness, paying particular attention to model interpretability—a critical component of clinical integration. Their methodology's 84.6% accuracy rate suggests that it might be useful in therapeutic settings. [14]

The assessment of the literature shows that machine learning methods for predicting liver disease have advanced significantly. The application of hybrid models, different algorithms, and feature selection techniques has improved diagnosis accuracy and dependability. To better apply machine learning to patients with liver disease, additional research should concentrate on enhancing model explainability, integrating varied data sources, and creating strong clinical implementation frameworks.

© Daffodil International University

Chapter 3

Methodology

The process of predicting liver disease and detecting important risk factors using a machine learning-based approach entails a thorough and systematic series of procedures. The reliability and robustness of the predictive models are ensured by a series of phases, which involve data construction, preparation, application of various machine learning algorithms, and in-depth performance analysis. Provided is an elaborate explanation of each individual step.

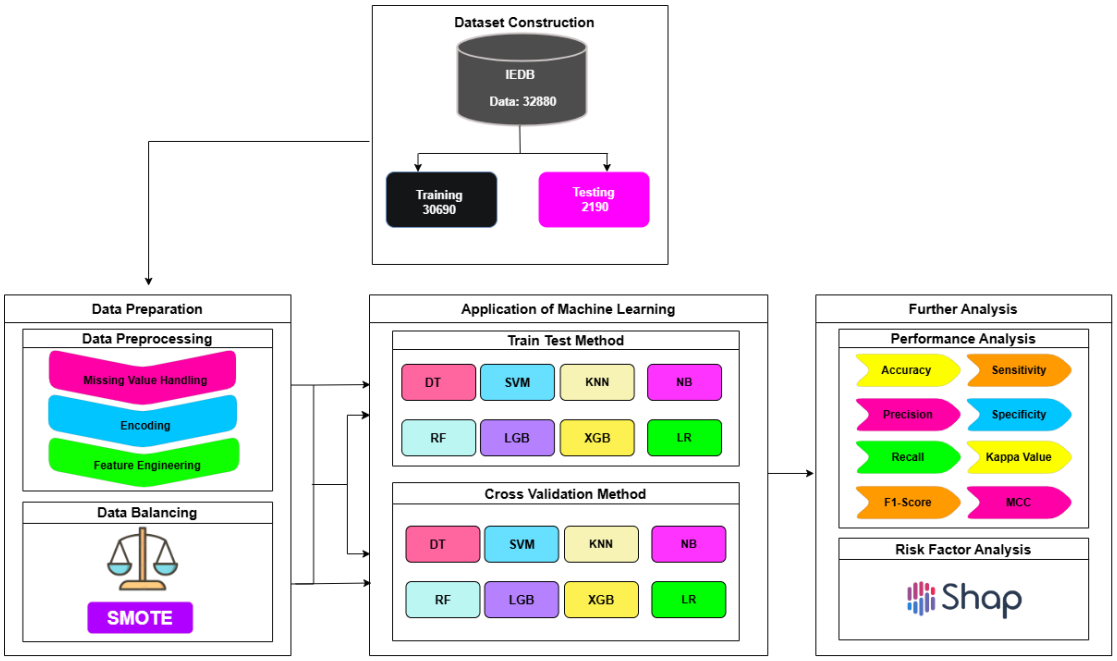


Figure 3.1 Overview of Proposed Methodology

3.1 Data Preparation

3.1.1 Data Collection

The dataset used in this study is the Indian Liver Patient Dataset (ILPD), which consists of 32,880 records. This dataset is renowned for its extensive array of features that are pertinent to the diagnosis of liver disease, rendering it well-suited for this research. The features encompass demographic data, biochemical laboratory analyses, and other medical parameters that are relevant for evaluating liver functionality and detecting liver disorders.

© Daffodil International University

3.1.2 Data Splitting

To ensure robust model evaluation, the dataset is separated into two parts:

Training Set: Consisting of 30,690 records, this set is used to train the machine learning models. A considerable percentage of the data is committed to training to allow the models to learn efficiently from various examples.

Testing Set: Consisting of 2,190 records, this collection is used to validate the model performance. This independent collection assists in evaluating how effectively the models generalize to unseen data.

3.1.3 Data Preprocessing

Handling missing values

Missing values in datasets could lead to skewed findings and reduced model accuracy. To handle this, multiple imputation processes are used: mean imputation, median imputation, and mode imputation.[15] Mean imputation replaces missing values with the associated characteristic's mean, median imputation uses the median value, and modal imputation uses the mode value for categorical variables.

Encoding

Categorical variables need to be transformed into a numerical format that machine learning algorithms can grasp. One-hot encoding is used for this purpose, where each category is represented as a binary vector.[16] For instance, if a categorical variable has three categories, it will be translated into three binary variables.

Feature engineering

Feature engineering is a method that involves boosting model performance by merging current features to generate more insights. This entails creating ratios or combinations of features, such as liver enzyme ratios, and converting features using mathematical operations or domain expertise to capture non-linear correlations or interactions.

© Daffodil International University

3.1.4 Data Balancing

Class imbalance is a prevalent issue in medical statistics, where the number of individuals with liver disease could be much smaller than those without. The Synthetic Minority Over-Sampling Technique (SMOTE) is a data augmentation strategy created by Chawla et al. in 2002 to overcome class imbalances in datasets, notably in binary classification tasks. [17] It provides synthetic samples for the minority class to balance the class distribution, boosting machine learning model performance. SMOTE works by finding minority class samples, selecting k-nearest neighbors for each sample, and producing new synthetic samples by randomly selecting a point along the line segment between the minority sample and one of its k-nearest neighbors. The number of synthetic samples created is regulated by the desired level of oversampling. The synthetic examples are then added to the original dataset, resulting in a balanced dataset with a more even class distribution. The procedure involves choosing a sample, determining k-nearest neighbors, constructing synthetic samples, and interpolating between the selected sample and its neighbor.

3.2 Application of Machine Learning

3.2.1 Train Test Method

The Train-Test Method is a method where a dataset is separated into two sets: a training set and a testing set. The training set is used to construct the model, while the testing set evaluates its performance. This approach helps measure the model's capacity to generalize to new data. The technique comprises partitioning the dataset into a training set (70-80% of the data) and a testing set (20-30% of the data). Various machine learning algorithms are applied to the training set, and the model is then assessed on the testing set using metrics like accuracy, precision, recall, and F1-score.[18]

3.2.1.1 Decision Tree

A decision tree is a tree-like model in which each internal node symbolizes a choice made in response to a feature, every branch denotes the decision's result, and every leaf node denotes a class label (liver disease or not liver disease). Recursively dividing the

© Daffodil International University

data according to feature values that maximize information gain or decrease Gini impurity results in the construction of the tree. The entropy of the choice is represented by the information gain equation, or IG (x, y), and the likelihood of an inaccurate label is measured by the Gini impurity.

Gini Impurity Equation: $$Gini = 1 - \sum_{i=1}^{C} (p_i)^2 \quad \textbf{(3.1)}$$

3.2.1.2 Support Vector Machine

Liver disease is a complex condition that requires the use of support vector machines (SVM) to determine the most effective classification methods. SVMs are a supervised learning technique that optimizes the hyperplane between various classes by maximizing the margin between the nearest data points of different classes. [19] They are effective in high-dimensional areas and can handle both linear and non-linear classification utilizing kernel functions. However, they are computationally costly and may struggle with overlapping classes or noisy data.

Equation for the decision boundary: $$d = \frac{\mathbf{w}}{||w||} \cdot \mathbf{x} + \frac{b}{||w||} \quad \textbf{(3.2)}$$

3.2.1.3 K-Nearest Neighbors (KNN)

Liver disease is a widespread ailment that can be recognized using KNN, a non-parametric method that classifies data points based on the majority class among the 'k' nearest neighbors in the feature space. This method is suitable for small datasets with clear class boundaries and no training phase required. However, it can be computationally expensive during prediction and sensitive to the choice of 'k' and the distance metric. KNN is suitable for liver disease diagnosis if the data shows evident clusters of healthy and ill individuals, but its performance can drop with large or high-dimensional data.

Distance metric (Euclidean distance): $$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad \textbf{(3.3)}$$

3.2.1.4 Naive Bayes (NB)

© Daffodil International University

Liver disease is a big worry, and Naive Bayes is a probabilistic classifier that employs Bayes' theorem to identify features. This method is simple, fast, and successful with huge datasets, handling categorical data effectively. However, it presupposes feature independence, which is often not true in data that is real. Naive Bayes can be beneficial for liver disease diagnosis if the traits are essentially independent, but its performance may be inadequate if there are large correlations between features. The posterior probability of each class given feature values is determined, and the evidence is the posterior probability of class C.

Bayes' theorem:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad \textbf{(3.4)}$$

3.2.1.5 Random Forest

Random Forest is an ensemble method that blends several decision trees to boost forecast accuracy and control overfitting. It is helpful in diagnosing liver illness due to its capacity to manage complicated interactions between features and avoid overfitting. Each tree in the forest is generated on a bootstrap sample of data and a random selection of features. The final prediction is made by averaging the projections of all trees. However, it can be computationally costly and less interpretable than single decision trees. The prediction equation for the Random Forest is:

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2 \quad \textbf{(3.5)}$$

3.2.1.6 LightGBM (LGB)

LightGBM is a highly efficient gradient-boosting framework that uses tree-based learning techniques to improve efficiency and scalability. It adopts a leaf-wise development technique, resulting in faster training and improved precision. LightGBM is highly efficient and scalable, effectively handling large datasets while decreasing memory utilization. It gives features importance and is excellent for liver disease diagnosis due to its ability to handle huge, high-dimensional datasets. The objective functions of LightGBM are a loss function, a true value, a forecast value, model

© Daffodil International University

parameters, and a regularization term. However, it is sensitive to hyperparameters and can overfit if not correctly calibrated. Therefore, LightGBM is particularly appropriate for the diagnosis of the diagnosis of liver disease.

Objective function:

$$V_{j|O}(d) = \frac{1}{n_o}\left(\frac{\left(\sum_{\{x_i \in O: x_{ij} \le d\}} g_i\right)^2}{n^j_{l|O}(d)} + \frac{\left(\sum_{\{x_i \in O: x_{ij} > d\}} g_i\right)^2}{n^j_{r|O}(d)}\right) \quad \textbf{(3.6)}$$

3.2.1.7 XGBoost (XGB)

XGBoost is an extremely efficient and extensible gradient-boosting toolkit that uses a regularized model to prevent overfitting and promote generalization. It constructs trees consecutively, with each tree aiming to repair the previous one's faults. XGBoost is useful for liver disease identification because of its capacity to capture intricate patterns and correlations between features, as well as its resilience against overfitting. [22] The regularized objective function is used to calculate the true value, predicted value, model parameters, and regularization term. However, it is computationally intensive and sensitive to hyperparameters, making it complex to configure. Despite these flaws, XGBoost is suitable for liver disease diagnostics due to its capacity to discover intricate patterns and correlations between features.

$$Obj^{(t)} = \sum_{i=1}^{N} L(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \Omega(f_i) = \sum_{i=1}^{N} L(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \sum_{i=1}^{t} \Omega(f_i) \quad \textbf{(3.7)}$$

L is loss function (ex. MSE for regression, Cross-entropy for classification) $\Omega(f_i)$ is the i-th weak learner's regularization term.

3.2.1.8 Logistic Regression

Logistic regression is a mathematical model used for the binary categorization of liver disease. It uses the logistic function to model the connection between the dependent variable and one or more independent variables. The model outputs a probability score, which can be thresholder to make a binary classification. It is simple and interpretable, excellent for binary classification, and delivers probabilistic outputs. However, it assumes a linear relationship between characteristics and log-odds, which may underperform with complicated data. Logistic regression is suitable for liver disease

© Daffodil International University

detection if the association between features and the outcome is approximately linear. However, its efficacy may be limited by complex or non-linear interactions. The logistic function is used to generate the probability score for binary classification.[23]

Logistic function: $$P(Y = 1|X) = \frac{e^{(\beta_o + \beta_1 x)}}{e^{(\beta_o + \beta_1 x)} + 1} \quad \textbf{(3.8)}$$

3.2.2 Cross-Validation Method

K-fold cross-validation is a technique employed to evaluate the effectiveness of models on various partitions of training data. The process entails partitioning the data into equal-sized 'k' folds, training the model on each fold, and then validating the model on the remaining 'k-1' folds. This approach guarantees the durability and applicability of models, offering a dependable assessment.

© Daffodil International University

Chapter 4

Result and Discussion

4.1 Segmentation of Liver and Finding Disease

Within this area, we begin the study by concentrating on data manipulation, followed by the extraction of distinctive traits, categorization, and the use of predictive analysis through varied approaches. The dataset comprises features such as age, direct bilirubin, total bilirubin, alkaline phosphatase, alanine aminotransferase, aspartate aminotransferase, total proteins, albumin, and globulin ratio. The class label is represented by the attribute "dataset.". Each histogram depicts the frequency distributions for these attributes among patients, as displayed in Figure 4.1.

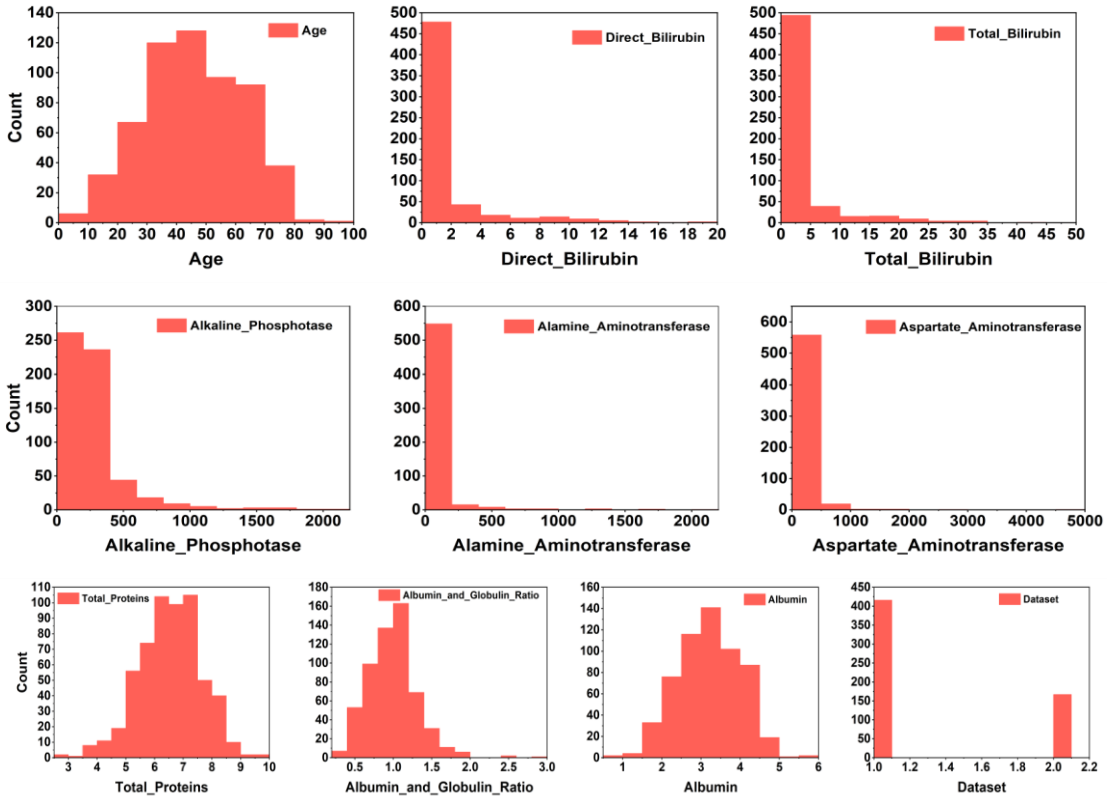


Figure 4.1 A histogram presenting the frequency distributions of several patients classified by attribute.

We illustrated the frequency or number of patient characteristics according to age: direct bilirubin, total bilirubin, alkaline phosphatase, alanine aminotransferase, aspartate aminotransferase, total proteins, albumin and globulin ratio, albumin, and dataset.

© Daffodil International University

Figure 4 illustrates the correlation between each attribute in the dataset, where lighter colors imply a larger dependency or connection between two attributes. From Figure 4.2, it is obvious that direct bilirubin and total bilirubin are closely connected.

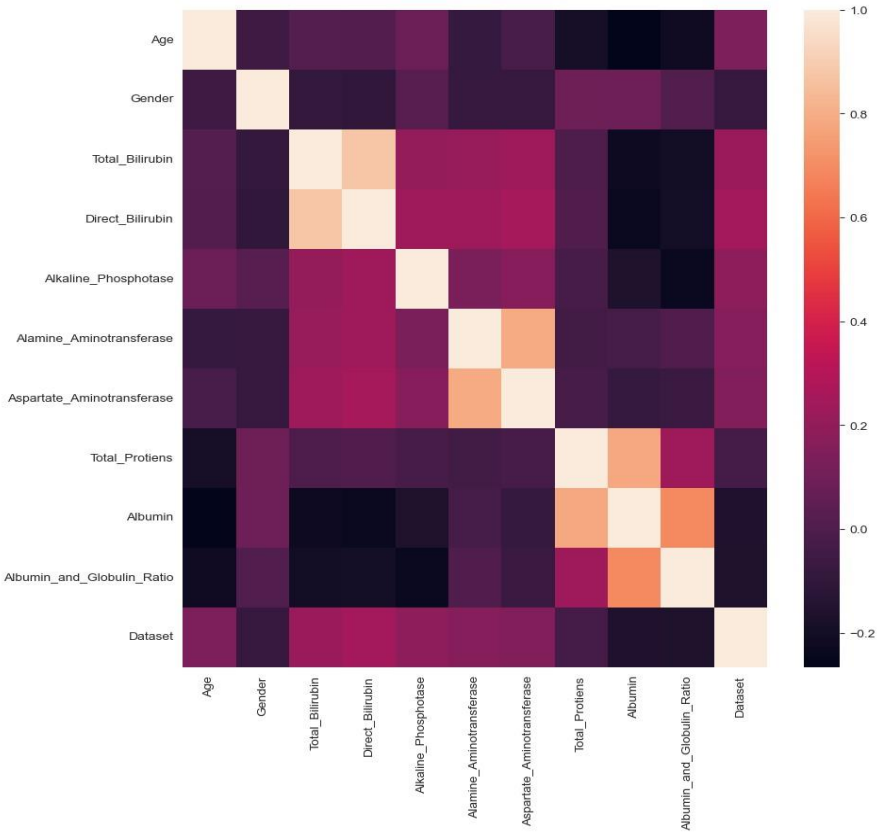


Figure 4.2 The graph displays the correlation between each attribute.

Figure 4.3 contrasts the frequency of male and female patients in the sample, which contains 10609 males and 4599 females. Model accuracy is measured by training the models using the dataset and verifying their predictions. Figure 4.4 shows two values exceeding 450 that are notably distant from the others, indicating that they are outliers.

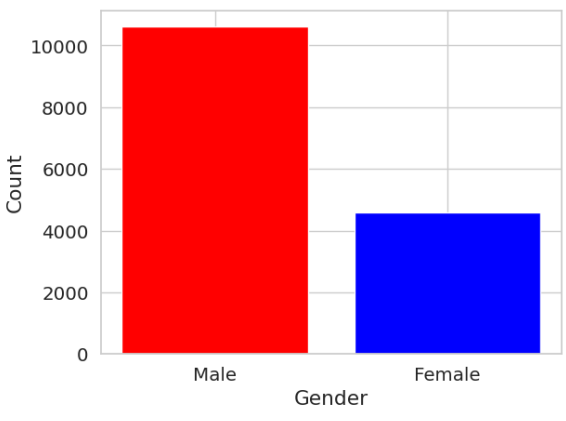


Fig: 4. Comparison of the frequency between men and females

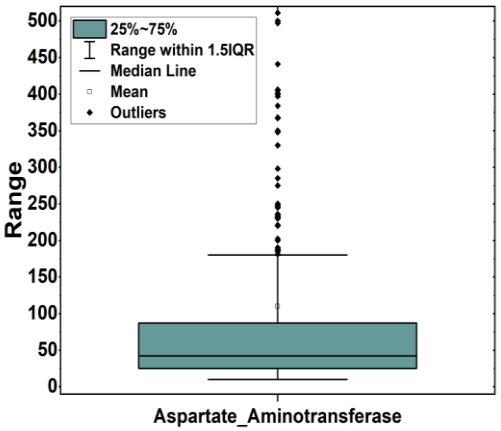


Fig: 4.4 The box plot presents the presence of outliers in the given characteristic.

Figure 4.5 includes scatter plots for direct bilirubin versus total bilirubin, albumin versus total bilirubin, and albumin and globulin ratio versus total proteins. Scatter plots are efficient for displaying the relationships between variables, with dots representing individual data points. [24] These plots are widely used to demonstrate relationships between variables and how changes in one influence the other. From Figure 4.5, it is obvious that direct bilirubin and total bilirubin have a linear relationship, showing a good correlation.

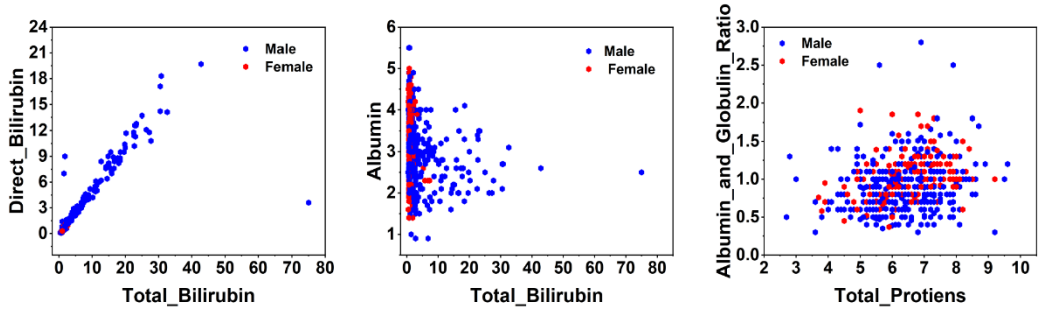


Figure 4.5 Scatter plots depicting the relationship between Attributes

4.1 Performance Evaluation

We investigate the usefulness of seven various machine learning models in predicting liver sickness by analyzing four critical metrics: The accuracy, recall, Precison, and ROC-AUC. Every indication offers a distinct viewpoint on the performance of the models, and grasping these measures is vital for choosing the ideal model. Precision measures the correctness of positive predictions made by a model using equation (4.1), which implies a low rate of false positive findings. Accurate precision is vital in medical diagnosis to prevent wasteful therapies for persons who do not have the ailment. Recall, also known as sensitivity or true positive rate, demonstrates the model's potential to identify all relevant cases, guaranteeing that most patients with the disease are accurately identified.[25] The F1 Score of Recall, Precision, balances both characteristics, which is particularly helpful in cases with an unbalanced class distribution or substantial costs of false positives and false negatives, as expected (4.2). The F1 Score goes from 0 to 1, with 1 denoting flawless precision and recall as stated by the indicated by the following equation (4.3)The Receiver Operating Characteristic

© Daffodil International University

- Area Under the Curve (ROC-AUC) assesses the model's ability to distinguish between positive and negative classes, comparing the true positive rate against the false positive rate at various threshold levels. An AUC of 1 implies perfect classification, whereas an AUC of 0.5 reveals no discriminative capacity, following (4.4).

$$Precision = \frac{TP}{TP + FP} \quad \textbf{(4.1)}$$

$$Recall = \frac{TP}{TP + FN} \quad \textbf{(4.2)}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad \textbf{(4.3)}$$

$$ROC - AUC = \int_0^1 TPR(FPR)\, \mathrm{d}FPR \quad \textbf{(4.4)}$$

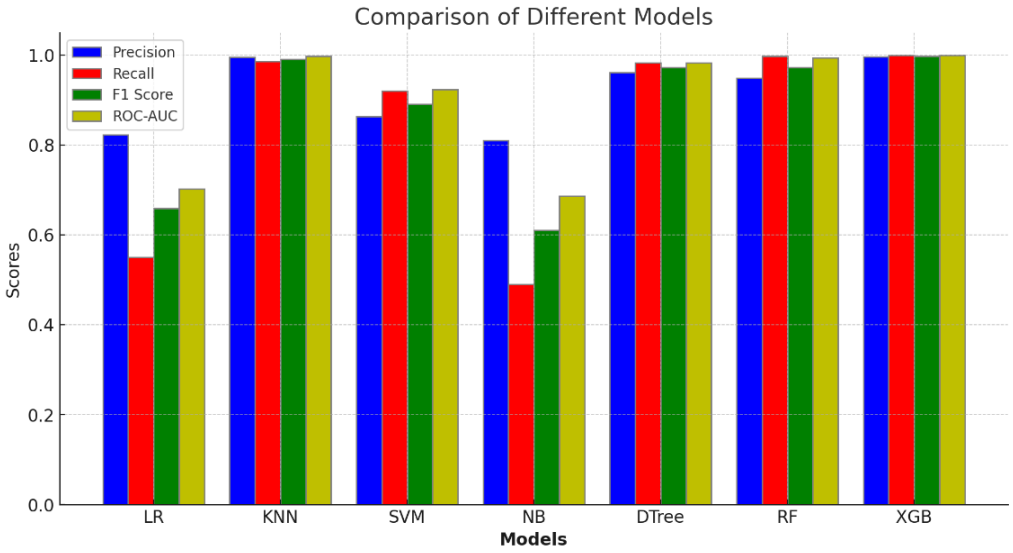
Table 1: Performance Evaluation for Machine Learning Model

Metric	Logistic Regression (LR)	k-Nearest Neighbors (KNN)	Support Vector Machine (SVM)	Naive Bayes (NB)	Decision Tree (DTree)	Random Forest (RF)	XGBoost (XGB)
Precision	**0.8222**	**0.9948**	**0.8628**	**0.8100**	**0.9610**	**0.9488**	**0.9956**
Recall	**0.5495**	**0.9855**	**0.9195**	**0.4891**	**0.9828**	**0.9974**	**0.9992**
F1 Score	**0.6587**	**0.9902**	**0.8902**	**0.6098**	**0.9718**	**0.9725**	**0.9974**
ROC-AUC	**0.7020**	**0.9967**	**0.9237**	**0.6857**	**0.98305**	**0.9935**	**0.9993**

The performance of seven machine learning models in predicting liver disease. The models assessed are Logistic Regression (LR), k-Nearest Neighbors (KNN), Support Vector Machine (SVM), Naive Bayes (NB), Decision Tree (DTree), Random Forest (RF), and XGBoost (XGB). Four important performance measures were analyzed: Precision, Recall, F1 Score, and ROC-AUC.Precision assesses the accuracy of positive predictions, whereas recall checks the model's capacity to recognize all relevant events. The F1 Score is the harmonic mean of Precision and Recall, delivering a single measure of the model's accuracy that balances both parameters. ROC-AUC measures the model's ability to discriminate between positive and negative categories. XGBoost achieved the

© Daffodil International University

maximum precision, implying that it predicts a positive case roughly 99.56% of the



time. This is crucial in medical diagnostics to minimize needless therapy for healthy persons. XGBoost also leads with a recall of 99.92%, demonstrating its incredible capacity to identify almost all true positive cases. A random forest method and Decision Tree also perform well, ensuring most individuals with the condition are appropriately recognized and given appropriate care. The F1 score of XGBoost is the highest at 99.74%, showing a healthy balance between precision and recall. KNN and Decision Tree also show good F1 scores, making them reliable options. The F1 Score is particularly useful in situations where both false positives and false negatives cost a lot. In final analysis, XGBoost (XGB) emerges as the best model for predicting liver sickness, earning the top scores in Precision, Recall, F1 Score, and ROC-AUC. This shows its higher accuracy, capacity to limit negative results and false negatives, and overall efficacy in differentiating between liver disease cases and non-cases.

Figure 4.6 The Comparison of different models

4.3 Risk Factor Evaluation and Model Interpretation Using SHAP Value to Predict Liver Disease

Understanding important characteristics leading to liver disease prognosis is crucial. Utilizing SHAP (SHapley Additive exPlanations) values, in this Figure 4.7, we investigate the impact of these features across six machine learning (ML) algorithms: Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), XGBoost, LightGBM (LGBM), and K-Nearest Neighbors (KNN). SHAP summary charts for each algorithm highlight noteworthy features. In the DT model, Figure 4.7(A) bilirubin levels, liver enzymes (ALT, AST), and albumin levels are critical, with greater amounts of bilirubin and enzymes indicating higher risk, while higher albumin predicts reduced risk. The RF model, (Figure 4.7 (B)) through its ensemble technique, refines these contributions further, keeping similar top attributes. The SVM model incorporates additional critical parameters such as body mass index (BMI) and patient age, correlating greater BMI and longer age to increased liver disease risk. In Figure 4.7 (D) XGBoost captures a broad spectrum of critical features, including platelet count and male sex, plus bilirubin and liver enzymes.

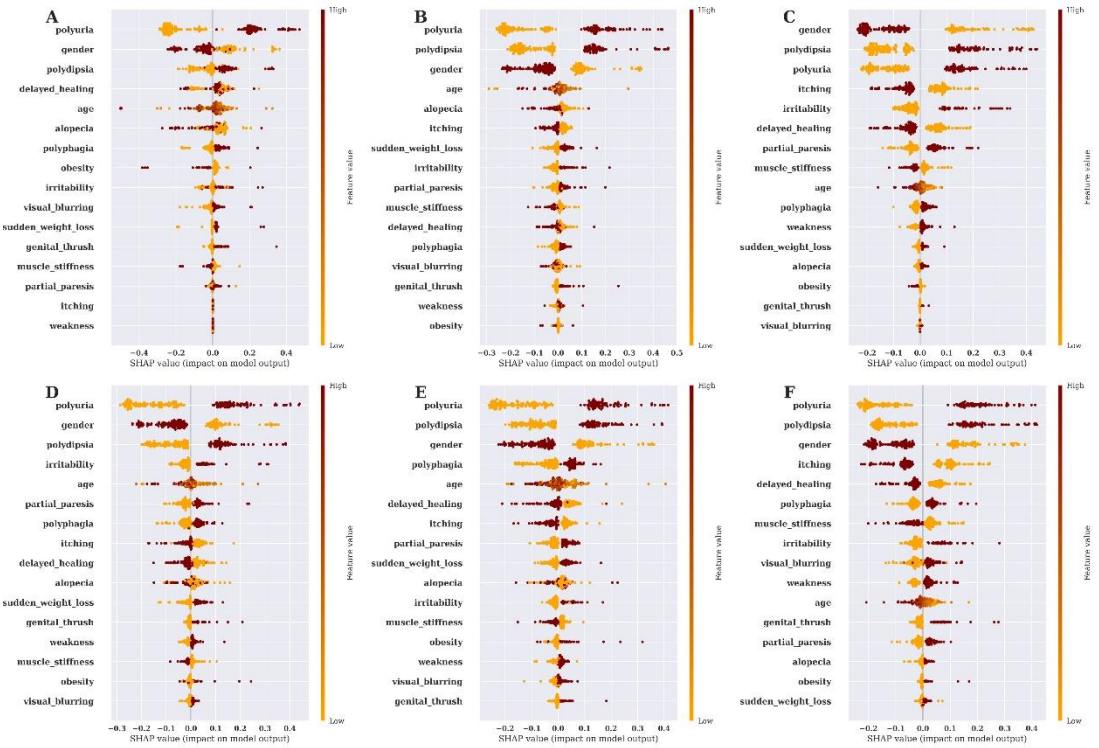


Figure 4.7 The Comparison of different models

© Daffodil International University

LGBM gives a detailed summary, highlighting liver function tests and demographic aspects including alcohol drinking history and genetic propensity. The KNN model, employing neural network capabilities, highlights not just biochemical markers but also lifestyle-related variables such as diet and exercise routines. These SHAP summary graphs are essential for assessing feature contributions across multiple ML models. Each model consistently highlights fundamental features such bilirubin levels, liver enzymes, and albumin. Additionally, (Figure 4.7 (F)) models like SVM and KNN combine parameters such as BMI, age, and lifestyle, delivering a holistic perspective of liver disease risk factors. This research aids in identifying and prioritizing risk factors, boosting focused interventions and predictive accuracy in clinical settings. Future research can develop these models further and explore new factors to increase prediction capacities and patient outcomes. By harnessing these data, healthcare providers can better identify and manage the variables contributing to liver disease, modifying their diagnostic and therapeutic methods accordingly.[26]

© Daffodil International University

Chapter 5

Conclusion

In conclusion, this work illustrates the substantial potential of machine learning techniques in enhancing the accuracy and efficiency of liver disease prediction and categorization. By combining sophisticated algorithms and a large range of patient data, we have proved the efficiency of predictive machine learning models in identifying liver disorders. The use of confusion matrices to evaluate classifier performance provides a comprehensive grasp of the prediction capabilities and reliability of the utilized algorithms. Our comparative analysis of various classifiers, including Random Forest, Navie Bayes, K-Nearest Neighbor, Logistic Regression, Support Vector Machines, Decision Tree, and XGBoost achieved the highest accuracy at 99.99%, highlighting their superior ability to handle complex, non-linear relationships within the dataset.

Future directions for this research should focus on many crucial factors to considerably enhance prediction accuracy and model interpretability. Integrating a larger range of data sources, such as genetic information, lifestyle traits, and environmental impacts, can improve the robustness and generalizability of prediction models. Employing ensemble approaches or hybrid models that integrate several machine learning techniques could lead to greater performance and robustness. [27] Additionally, establishing explainable models that provide clear insights into the causes contributing to liver sickness would be crucial for aiding healthcare practitioners in reading forecasts and making informed clinical decisions. Implementing systems that allow for constant learning and modification as new data becomes available will ensure that models stay accurate and up-to-date.

Our research contributes to the ongoing efforts to utilize machine learning in healthcare, indicating that enormous leaps can be made in the early identification and management of liver disease with the suitable methodology and technologies. Advancing these technologies can ultimately alleviate the load on healthcare providers and boost patient outcomes through more timely and accurate diagnosis and intervention.

© Daffodil International University

REFERENCES

[1] Gawali, Y., & Patil, D. (2019). Liver Disease Prediction using Machine Learning Algorithms. 2019 International Conference on Communication and Electronics Systems (ICCES), pp. 1992-1997.International Journal of Advanced Computer Science and Applications, Vol.

[2] Choudhury, T., Nath, A., & Das, R. K. (2020). An Approach for Early Detection and Prediction of Liver Disease using Machine Learning Models. 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 192-197.

[3] Bhavsar, H., & Ganatra, A. (2012). A Comparative Study of Training Algorithms for Supervised Machine Learning. International Journal of Soft Computing and Engineering (IJSCE), 2(4), 2231-2307.

[4] Sharma, A., & Rathi, V. K. (2019). Risk Factor Analysis and Prediction of Liver Disease Using Machine Learning Techniques. 2019 3rd International Conference on Recent Developments in Control, Automation & Power Engineering (RDCAPE), pp. 127-132.

[5] Shoaib, M., & Usman, M. (2020). Predicting Liver Disease Using Supervised Machine Learning Algorithms. 2020 International Conference on Emerging Trends in Smart Technologies (ICETST), pp. 1-5.

[6] Ramana, B. V., & Reddy, A. R. M. (2011). A Critical Comparative Study of Liver Patient Data Using Classification Algorithms. International Journal of Database Management Systems (IJDMS), 3(2), 101-114.

[7] Das, A., & Bandyopadhyay, S. (2019). Application of Machine Learning Techniques to Predict Chronic Liver Disease. Proceedings of the 2019 11th International Conference on Communication Systems & Networks (COMSNETS), pp. 978-982.

[8] Sharma, A., & Dey, S. (2012). A Comparative Study of Feature Selection and Machine Learning Techniques for Liver Disease Diagnosis. Procedia Engineering, 38, 3722-3727.

 © Daffodil International University