Liver Disease Risk Prediction: A Machine Learning Approach

BY

Md Yasin Arafat Shuvo

[redacted]

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Dr. S. M. Aminul Haque

[redacted]

Co-Supervised By

Md. Sazzadur Ahamed

[redacted]

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2024

APPROVAL

This Project titled **"Liver Disease Risk Prediction: A Machine Learning Approach"**, submitted by Md Yasin Arafat Shuvo, ID No: 201-15-13706 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 24 January 2024.

BOARD OF EXAMINERS

[redacted]

Chairman

Dr. Sheak Rashed Haider Noori (SRH)
Professor & Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

[redacted]

Internal Examiner

Raja Tariqul Hasan Tusher(THT)
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

[redacted]

Internal Examiner

Mayen Uddin Mojumdar(MUM)
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

[redacted]

External Examiner

Dr. Md. Arshad Ali (DAA)
Professor
Department of Computer Science and Engineering
Hajee Mohammad Danesh Science and Technology
University

DECLARATION

I hereby declare that, this project has been done by myself under the supervision of **Dr. S. M. Aminul Haque, Professor & Associate Head, Department of CSE** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Sup[redacted]

Dr. S. M. Aminul Haque
Professor & Associate Head
Department of CSE
Daffodil International University

Co-Supervised by:

[redacted]

Md. Sazzadur Ahamed
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:

[redacted]

Md Yasin Arafat Shuvo
ID: 201-15-13706
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, I express our heartiest thanks and gratefulness to almighty God for His divine blessing makes me possible to complete the final year project successfully.

I really grateful and wish my profound my indebtedness to **Dr. S. M. Aminul Haque, Professor & Associate Head**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of my supervisor in the field of “Machine Learning, Artificial Intelligence, Data Mining, Distributed and High-Performance Computing” to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

I would like to express my heartiest gratitude to **Dr. Sheak Rashed Haider Noori, Professor and Head,** Department of CSE, for his kind help to finish my project and also to other faculty member and the staff of CSE department of Daffodil International University.

I would like to thank my entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

Liver disease is a serious worldwide health problem with high mortality rates, emphasizing the significance of early detection and treatment for better patient outcomes. Cirrhosis, fatty liver disease, fibrosis, viral hepatitis, and liver cancer are the most common liver conditions. In order to improve the prediction of liver disease risk, machine learning (ML) has shown great promise. Using a large patient dataset, the main objective of this research is to create and assess machine-learning models that predict the risk of liver disease. We used and compared a number of machine learning (ML) algorithms, such as logistic regression, support vector machines (SVMs), decision trees, random forests, k-neighbors, and gradient-boosting models. Notably, the outcomes demonstrate how much more accurate the gradient boosting model is than other machine learning techniques. Categories including "blood donor," "suspect blood donor," "hepatitis," "fibrosis," and "cirrhosis" were among the important characteristics considered in this predictive model. We were able to improve the gradient boosting model's predicted accuracy by carefully examining these variables. The results of this investigation show that the gradient boosting model performed substantially better than the other models, with 94% accuracy and f1-score, 95% precision, and 94% recall. These results highlight the enormous potential of advanced machine learning methods to revolutionize the field of risk detection for liver disease.

TABLE OF CONTENTS

CONTENTS **PAGE**

CHAPTER

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
Introduction

1.1 Introduction

Lying behind the rib cage in the upper right abdomen is the liver, the biggest organ in the body. The liver is an organ that seems like a wedge and has two lobes that are irregularly sized and shaped. The liver is dark reddish brown in hue. A human liver typically has dimensions of 15 centimeters (6 inches) by 1.5 kilograms (3.3 pounds). The typical reference range for women is 600-1,770 g (1.32–3.90 lb.), and for men it is 970-1,860 g (2.14–4.10 lb.). This suggests that there is a notable variation in individual sizes. In addition to being the heaviest internal organ, it is the biggest gland in the entirety of the human body. It is situated directly above the stomach and below the diaphragm in the upper right quadrant of the abdominal cavity, covering the gallbladder. The portal vein and the hepatic artery are two vital arteries that supply blood to the liver. Using the portal vein, blood from the pancreas, spleen, and entire gastrointestinal tract is received by the hepatic artery, which transports oxygen-rich blood from the aorta via the celiac trunk. These blood vessels divide to form the liver sinusoids, which are tiny capillaries that eventually split off to form hepatic lobules. It is an essential organ that performs numerous vital functions, including the synthesis of proteins, digestion, metabolism, detoxification, glycogen, vitamin, and mineral storage.

1.2 Motivation

Liver disease has grown into a bigger public health concern, with a reported 2-3 million deaths from it happening every year globally. Cirrhosis, fatty liver disease, fibrosis, viral hepatitis, and liver cancer are the most widespread forms of liver disease. Damage to the liver can lead to various diseases, including cirrhosis, liver cancer, and liver failure. Liver damage can lead to different medical issues such as jaundice, hepatic encephalopathy (HE), and ascites. If liver problems are detected early on, the patient's chances of survival are better. Successful therapy and prevention of complications can be helped by an accurate early detection and prognosis of liver disease. We can preserve the health of our liver by acquiring hepatitis vaccination, abstaining from excessive alcohol use, maintaining a healthy weight, and following a balanced diet. However, because there might not be any symptoms in the early stages, liver disease can be challenging to diagnose. Because of this, many liver disease patients do not receive a diagnosis until their condition has advanced to a more serious stage.

1.3 Rationale of the Study

Early on in progression, liver diseases are frequently asymptomatic, making diagnosis challenging to determine who is most likely to develop liver disease. Numerous variables, such as lifestyle, medical history, demographics, and laboratory testing, can be used to do this. Clinicians can focus interventions and monitoring to stop or postpone the onset of liver disease by identifying high-risk individuals.to create and verify models for risk prediction. The chance that a person will develop liver disease over a specific time period can be determined through these models. Making decisions regarding treatment, monitoring, and screening can be aided by this information. to increase understanding of the liver disease risk factors. Through the identification of contributing factors to liver disease, researchers can devise more efficacious strategies for prevention and treatment.

1.4 Research Questions

1. Which risk factors for liver disease are especially noteworthy?
2. What steps can we take to enhance liver disease earlier detection?
3. How can we create more precise and trustworthy techniques to estimate the risk of liver disease?
4. How can liver disease prediction be made better by artificial intelligence and machine learning?
5. How can we create models of specific liver disease risk testing?
6. How can liver disease prediction be implemented into clinical practice?

1.5 Expected Output

The expected result of liver disease risk prediction is enhanced liver disease detection and prevention. By determining people who are at a higher risk of liver disease, medical professionals may focus on treatments and keep track to stop or slow the progression of the disease. For example, individuals who have liver disease who are particularly susceptible to developing the disease might get greater amounts of treatment and surveillance. Utilizing risk prediction models can help clinicians make better decisions regarding treatment, monitoring, and screening. For example, a model for risk prediction can be used to detect which patients should be examined for liver disease and at what frequency. A potentially useful technique for predicting the risk of liver disease is machine learning (ML). Clinicians can identify patients who are at high risk for liver disease and take early action to stop or slow the disease's progression by using these tools. Large patient data sets can be analyzed using machine learning algorithms to find trends linked to liver disease. Models that forecast a new patient's risk of liver disease can then be created using these patterns. In this work, we used a sizable patient data set to create and assess machine learning models for predicting the risk of liver disease. The effectiveness of several machine learning (ML) algorithms, such as logistic regression, K neighbors, decision trees, random forests, support vector machines (SVMs), and gradient boosting models, was compared. We also identify the most important features for liver disease risk prediction.

1.6 Project Management and Finance

For any liver disease risk prediction project to be effective, project management and financing are essential components.

The process of arranging, planning, and managing resources to accomplish a certain objective is known as project management. The effective completion of the liver disease risk prediction project can be ensured through efficient project management. Project managers can assist in keeping the project on track and preventing expensive delays or scope creep by precisely defining the objectives, scope, and timeline. Project managers may assist in identifying and addressing such issues early on, before they have a substantial influence on the project, by keeping an eye on the status of the work and making necessary adjustments. Additionally, by maintaining clear paths of communication with every person involved, project managers can guarantee that everyone is on the same page regarding the project's objectives.

Managing money is the process of finance. To guarantee that the liver disease risk prediction project is finished within budget, effective financial management might be helpful. Financial managers can assist in determining and projecting expenses by creating a complete project budget. Financial managers can keep an eye on project expenses and identify any possible cost overruns by keeping track of project expenses. Financial managers can guarantee that a project has the resources necessary to finish it by locating and securing finance. Financial managers may assist in protecting the project from unexpected financial losses by controlling financial risks.

1.7 Report Layout

Chapter 1: The research introduction, aims, and important research questions are presented.

Chapter 2: Brief summaries of the literature review are presented.

Chapter 3: The proposed methodology is described in depth in.

Chapter 4: The experimental results of the paper are explained and explored.

Chapter 5: Describes about the sustainability and social impact.

Chapter 6: Ends the current research and provides a plan for future effort.

CHAPTER 2

Background

2.1 Preliminaries

1. Risk Factor

Any field of study that builds the prospect of acquiring a disease is regarded as an indicator of risk. There are merely a pair of hazards for liver disease: modifiable and non-modifiable. Risk elements that can be altered consist of drinking alcohol, malnutrition, and eating habits. Risk variables that are unchangeable, like age, sex, and family history, are recognized as non-modifiable risk factors.

2. Risk Prediction Model

A computational framework called a risk prediction model makes use of hazards to forecast a person's risk of contracting a disease. Large patient data datasets are commonly utilized in the creation of models that predict risks for liver disease. The purpose of training the computational models is to identify data patterns associated with liver disease. Once prepared, the models can be utilized for predicting a new patient's risk of liver disease.

3. Machine Learning (ML)

The use of artificial intelligence (AI) in the form of machine learning (ML) allows machines to learn without any direct programming. By investigating significant patient information sets, machine learning (ML) algorithms can be used to develop risk prediction models for liver disease.

4. External Validation

Determining a risk prediction model through a collection of data that wasn't utilized to create the model is referred to as external validation. To be able to ensure that the model may be utilized in new populations, external validation is crucial.

2.2 Related Works

The study by Bhupati, Deepika & Tan (2022) developed a machine learning model to predict the liver disease detection using data from the Indian Liver Patient records. The model was able to predict liver disease with an accuracy of 92.1%. The model used a variety of features, including demographic data, laboratory results, clinical data, and viral load data.

The study by Dritsas, E.; Trigka, M. (2023) developed a machine learning model to predict the risk of liver disease using data from the Indian Liver Patient records. The model was able to predict liver disease with an accuracy of 80.10%. The model used a variety of features, including demographic data, laboratory results, clinical data, and viral load data.

The study by Gajendran. (2020) developed a machine learning model to predict the risk of liver disease using data from the Indian Liver Patient records. The model was able to predict liver disease with an accuracy of 75.30%. The model used a variety of features, including demographic data, laboratory results, clinical data, and lifestyle data.

The study by C. Geetha (2021) developed a machine learning model to predict the risk of liver disease in patients with evaluation-based approach using data from the Indian Liver Patient records. The model was able to predict liver disease with an accuracy of 75.04%. The model used a variety of features, including demographic data, laboratory results, clinical data, and medication data.

The study by Rahman, A.K.M Sazzadur et al. (2019) developed a machine learning model to predict the risk of liver disease. The model was able to predict HCC with an accuracy of 75%. The model used a variety of features, including demographic data, laboratory results, imaging data, and clinical data.

The study by Ajay & Irfan (2018) developed a machine learning model to predict the risk of liver disease. The model was able to predict liver fibrosis with an accuracy of 75%. The model used a variety of features, including demographic data, laboratory results, and imaging data.

2.3 Comparative Analysis and Summary

All 6 studies used datasets of liver disease patients to develop and evaluate their models. They also used a variety of machine learning algorithms, including voting, logistic regression, SVM and MAMFFN.

The overall performance of the models was very good, with all 6 studies achieving accuracies of over 92.1%. The models also had high specificities and sensitivities, suggesting that they were able to accurately identify both patients with and without liver

Table 2.3 Comparative analysis of related works

Study	Dataset	Model	Accuracy	Specificity	Sensitivity	Precision
Bhupati, Deepika & Tan (2022)	ILPD	Autoencoders	92.1%	98.7%	87.65%	92.1%
Dritsas, E.; Trigka, M. (2023)	ILPD	Voting	80.1%	80.4%	80.1%	80.4%
Gajendran. (2020)	ILPD	MAMFFN	75.3%	74.67%	73.96%	71.36%
C. Geetha (2021)	ILPD	SVM	75.04%	71.11%	79%	77.09%
A.K.M Sazzadur et al. (2019)	ILDP	LR	75%	47%	78%	91%
Ajay & Irfan (2018)	ILDP	K-NN	73.97%	31.7%	90.4%	90%

disease. The highest overall performance was achieved by the Autoencoders models developed by Bhupati, Deepika & Tan (2022), which achieved accuracies, specificities and sensitivities of 92.1%, 98.7% and 87.76% respectively. However, the results of these studies suggest that machine learning can be used to develop accurate and reliable risk prediction models for liver disease. These models could be used to improve the early detection and intervention of liver disease, leading to better patient outcomes.

2.4 Scope of the Problem

Liver disease is a major global health problem, affecting over 1 billion people worldwide. It is the 11th leading cause of death globally, and the leading cause of death among adults in the United States aged 25-44. Liver disease can lead to a number of complications, including:

1. Liver failure: When the liver can no longer operate normally, it is a dangerous condition known as liver failure. Hepatic encephalopathy, ascites, jaundice, and bleeding are just a few of the health issues that can result from liver failure. A liver transplant is necessary for people with liver failure, a condition that is fatal.

2. Hepatocellular carcinoma (HCC): Primary liver cancer contains HCC. The most widespread sort of liver cancer, HCC stands sixth around the globe in terms of deaths due to cancer. With a poor prognosis, HCC is a difficult cancer to treat.
3. Portal hypertension: Blood pressure rising in the portal vein, and this is the vein that transports blood from the intestines to the liver, is referred to as portal hypertension. A wide range of issues, among them variceal hemorrhage, ascites, and splenomegaly, may occur in portal hypertension.

2.5 Challenges

1. Lack of large, well-annotated datasets: Developing accurate machine learning models requires large datasets of labeled data. However, there is a lack of large, well-annotated datasets of patients with liver disease. This is due to the fact that liver disease is a complex disease with a variety of different causes and presentations. Additionally, liver disease can be difficult to diagnose, and may require invasive procedures such as liver biopsy.
2. Data heterogeneity: The data used to develop and evaluate machine learning models for liver disease risk prediction is often heterogeneous, meaning that it comes from a variety of different sources and may have different formats. This can make it difficult to develop models that are generalizable to new populations and clinical settings.
3. Model interpretability: Machine learning models may be intricate and challenging to comprehend. Because of this, clinicians may find it challenging to trust and apply the models in their work.
4. Ethical considerations: The development and use of machine learning models for liver disease risk prediction raises a number of ethical considerations. For example, it is important to ensure that the models are used fairly and equitably, and that they do not perpetuate existing health disparities.

CHAPTER 3

Research Methodology

3.1 Research Subject and Instrumentation

We design our model into seven major steps. First of all, we put data acquisition then comes data profiling, third and fourth step is data scaling and features selection, after that classification and at last comes results step and data visualization. These steps are the basic structure of our proposed model. The full structure of our work is given below:

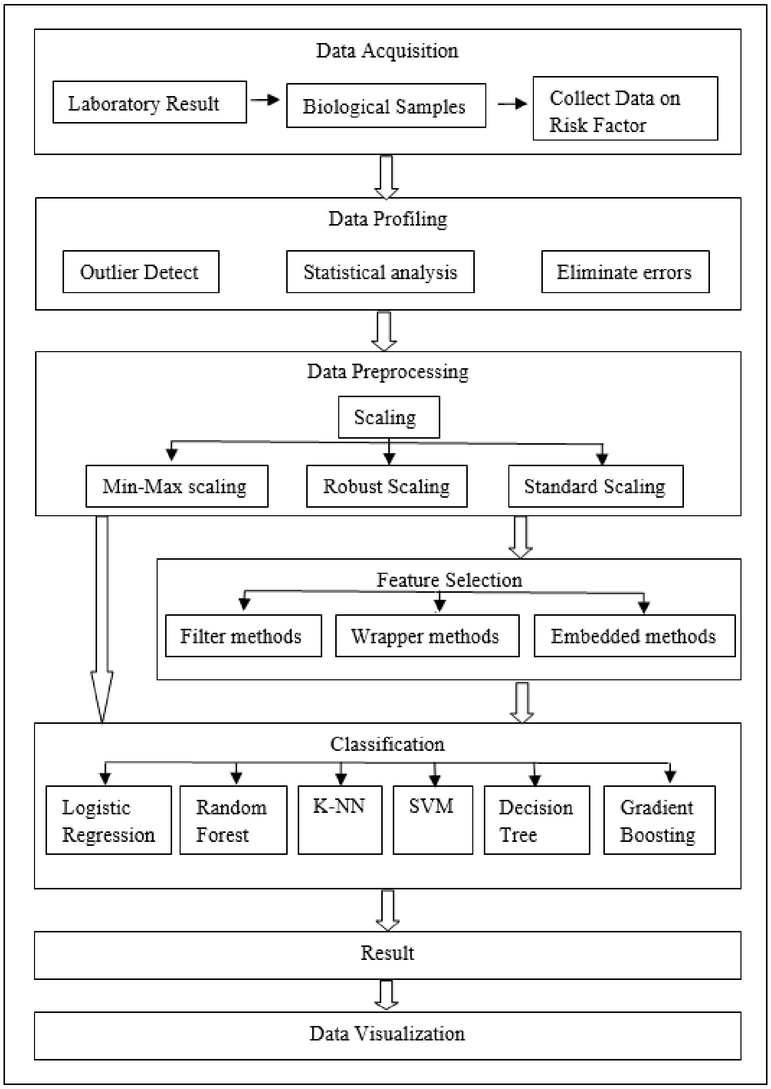


Figure 3.1 Methodology

3.2 Data Collection Procedure/Dataset Utilized

The dataset used in our study included demographic and laboratory data from Hepatitis C patients as well as blood donors. The data was obtained from UCI Machine Learning Repository.

Table 3.2 Dataset Description.

Feature	Type	Description
Sex	categorical	This feature illustrates the participant’s sex.
Category	categorical	This feature illustrates the diagnosis values
Age	numeric	The age range of the participants in 19-77 years.
ALB (Albumin Blood Test)	numeric	This feature captures the participant’s Albumin Blood Test in range 14.9-82.2
ALP (Alkaline phosphatase)	numeric	This feature captures the participant’s Alkaline phosphatase in range 11.3-416.6
ALT (Alanine Transaminase)	numeric	This feature captures the participant’s Alanine Transaminase in range 0.90-325.3
AST (Aspartate Transaminase)	numeric	This feature captures the participant’s Aspartate Transaminase in range 10.6-324
BIL (Bilirubin)	numeric	This feature captures the participant’s Bilirubin in range 0.8-254
CHE (Acetylcholinesterase)	numeric	This feature captures the participant’s Acetylcholinesterase in range 1.42-16.41
CHOL (Cholesterol)	numeric	This feature captures the participant’s Cholesterol in range 1.43-9,67
CREA (Creatinine)	numeric	This feature captures the participant’s Creatinine in range 8-1079.1
GGT (Gamma-Glutamyl Transferase)	numeric	This feature captures the participant’s Gamma-Glutamyl Transferase in range 4.5-650.9
PROT (Proteins)	numeric	This feature captures the participant’s Proteins in range 44.8-90

3.3 Statistical Analysis

There are 615 instances and 14 features in our dataset. Among them, 12 features are numerical and 2 are categorical features denoted by ‘Category’ and ‘Sex’. ‘Category’ includes five categorical values that are ‘0=Blood Donor’, ‘0s=suspect Blood Donor’, ‘1=Hepatitis’, ‘2=Fibrosis’, and ‘3=Cirrhosis’. ‘Sex’ includes two categorical values that are ‘Male’ and ‘Female’. ‘ALB (Albumin Blood Test)’, ‘ALP (Alkaline phosphatase)’, ‘ALT (Alanine Transaminase)’, ‘AST (Aspartate Transaminase)’, ‘BIL (Bilirubin)’, ‘CHE (Acetylcholinesterase)’, ‘CHOL (Cholesterol)’, ‘CREA (Creatinine)’, ‘GGT (Gamma-Glutamyl Transferase’ and ‘PROT (Proteins)’ are the laboratory data.

3.4 Proposed Methodology/Applied Mechanism

1. Data Preprocessing

The most important phase in the preparation of data is preprocessing. By appropriately scaling and converting the data set, data preprocessing facilitates the training and testing process. Machine learning algorithms must be preprocessed before they can be trained. Preprocessing brings features into balance and removes outliers. The majority of the time, scaling and normalization algorithms are used to improve learning.

i. Data Normalization

Reforming the nominal attributes in the 0–1 range involves using data normalization. To reduce the training process's sensitivity to feature size, data normalization is also utilized in machine learning. In essence, it makes our model more accurate by assisting in its convergence to better weights. When features are more compatible with one another, the model can predict outcomes more accurately. This is known as standardization.

ii. Data Scaling

Data scaling involves converting data into a predetermined range, like 0-100 or 0-1, so that it fits. Scaling data is important when using measuring tools that depend on the amount of distance separate data points, like support vector machines (SVM), k-nearest neighbors, or KNN. Any numeric feature that changes by '1' is given equal weight when using these algorithms.

a) Min Max Scaling

The min-max normalization technique is one of the most effective for normalizing data. The maximum value for each feature is converted to a 1, the minimum value to a 0, and all other values are altered to a decimal between 0 and 1. The following formula is typically used to achieve Min-Max scaling:

$$Msc = (M - Mmin)/(Mmax - Mmin)$$

Where M is the cell's initial value, Mmin is the column's minimum value, Mmax is the column's maximum value, and Msc is our new value. This scaler performs better in situations where the regular scaler isn't as effective. When the standard deviation is extremely low or the variance is non-Gaussian, the min-max scaler performs best.

b) Robust Scaling

The Robust Scaler takes a similar approach to the Min-Max scaler, but instead of using the min-max because the data points are more robust, it uses the confidence interval. It adheres to the formula for every feature in this way:

$$Xr = xi – Q1(x)/Q3(x) – Q1(x)$$

This implies that less data is used for scaling, because where there are data points, it is more appropriate

c) Standard Scaling

Standardization is a technique of rescaling which focuses on the nature of the variables on value 0 and the standard deviation on value 1. Together, to summarize a normal distribution, the mean and the standard deviation can be used. It requires that, before scaling, the mean and standard deviation of the values for each column be known. For a set of numbers, the mean determines the middle or main pattern. The column average is determined as the sum of all column values divided by the total number of variables.

$$mean = sum(values)/totalvalues$$

The average distribution of values from the mean is defined by the standard deviation. The sum of the square difference between each value and the mean can be determined as a square root and divided by the number of values minus 1.

$$standarddeviation = sqrt((valuei - mean)\ 2/(totalvalues - 1))$$

We can conveniently standardize the values in each column once the summary statistics are computed. To standardize a given value, the calculation is as follows:

$$\mathit{standardizedvalue} = (\mathit{value} - \mathit{mean})/\mathit{stdev}$$

iii. Feature Selection

Reducing the quantity of input data when developing a predictive model is known as feature selection. Reducing the sample size is desirable in order to reduce the modeling's computational challenge and, in particular circumstances, improve the model's efficiency. The use of statistics to determine the correlation between each input variable and the target variable and select the input variables with a high correlation with the target feature are examples of statistically based feature selection techniques. These methods can be efficient and fast; however, the selection of statistical measures is dependent on the type of data for both the input and output variables.

a) Filter Methods

Filter methods rank each feature individually based on some statistical measure, such as correlation, information gain, or chi-squared. The top-ranked features are then selected for the final model. Filter methods are computationally efficient and can be used on large datasets, but they can be less accurate than other feature selection methods because they do not consider the interactions between features.

- Correlation coefficient

The correlation coefficient quantifies the strength of a two-variable linear relationship. It is useful for identifying features that are substantially related with the target variable.

- Chi-squared test

The chi-squared test is a statistical test used to determine the relationship between two category variables. It can be used to select features that are correlated with the target variable.

- Fisher's score

The statistical significance of the difference between the means of two groups is indicated by the Fisher's score. It can be used to select features that are able to discriminate between the different classes of the target variable.

- Variance threshold

The variance threshold removes features that have low variance. This can be useful for removing irrelevant or noisy features.

- Mean absolute difference

The mean absolute difference measures the average difference between the values of two features. It can be used to select features that are different between the different classes of the target variable.

- Dispersion ratio

The dispersion ratio measures the ratio of the between-class variance to the within-class variance for a feature. It can be used to select features that are discriminative between the different classes of the target variable.

b) Wrapper methods

Wrapper methods are a subset of feature selection techniques which assess the performance of various feature subsets using a machine learning algorithm. Then, the feature subset with the best performance is chosen. Although they require more computing power, wrapper techniques are more precise than filter techniques. This is due to the fact that wrapper methods require the machine learning model to be trained and assessed on numerous feature subsets.

- Recursive feature elimination (RFE)

RFE is a greedy algorithm that begins with every feature and recursively removes the least significant feature until a predetermined number of features are left. Every feature's significance is determined by using the subset of features to train a machine learning model and then assessing the model's output.

- Sequential forward selection (SFS)

SFS is a greedy algorithm that adds the most significant feature one after another until a predetermined number of features are added. It begins with no features. By using the current subset of features to train a machine learning model and adding the feature that yields the largest performance improvement, the significance of each feature is determined.

- Exhaustive search

Every conceivable subset of features is assessed in an exhaustive search. Next, the subset of features that performs the best on the training set is chosen. The most accurate feature selection technique is exhaustive search, but it also requires the most computing power.

c) Embedded methods

Machine learning algorithms themselves incorporate embedded feature selection techniques. They choose the most significant features based on an internal evaluation metric built into the algorithm. Although embedded methods are more efficient and can be used with more complex machine learning algorithms, they are generally less accurate than wrapper methods.

- Lasso regression

Lasso regression is a type of logistic regression that performs L1 regularization. L1 regularization penalizes the coefficients of the model, which forces the model to select fewer features.

- Ridge regression

Ridge regression is a type of logistic regression that performs L2 regularization. L2 regularization also penalizes the coefficients of the model, but to a lesser extent than L1 regularization.

- Decision trees

Decision trees are a sort of machine learning method that can be used for classification as well as regression. Decision trees learn a set of rules for predicting the target variable. The rules are discovered by iteratively splitting the data into smaller and smaller groups based on feature values.

- Random forests

Random forests are an ensemble learning method that predicts using numerous decision trees. Random forests are more resistant to overfitting than individual decision trees.

2. Data Classification

Data classification marks the data according to the type, sensitivity and importance of the entity whether it is changed, stolen or lost. It helps a company understand the importance of its data, decides whether the data is at risk, and introduces risk reduction controls.

i. Logistic Regression

One statistical model that is frequently applied to classification tasks is logistic regression. This algorithm learns from a dataset of labeled examples because it is supervised learning.

Predicting the likelihood of an event occurring given a set of input variables is the aim of logistic regression. When one or more predictor variables, like whether a patient has a specific disease, are present, logistic regression is frequently used to predict binary outcomes (0 or 1, Yes or No, True or False).

ii. Random Forest

A supervised machine learning algorithm called random forest can be applied to tasks involving regression and classification. It is predicated on the idea of ensemble learning, which is the addition of several classifiers to enhance the model's overall performance. During training, a large number of decision trees are built by random forests. Random subsets of both the data and the features are used to train each decision tree. By doing this, the model's generalization performance is enhanced and overfitting is lessened.

iii. K-Nearest Neighbors (K-NN)

K-NN, or K-Nearest Neighbors, is a machine learning technique that is simple to use and excellent for regression and classification tasks. After selecting the K data points that are most similar to the new data point, it guesses the class or value of the new data point based on the classes or values of the K nearest neighbors. Because K-NN is a non-parametric technique, it makes no assumptions about the underlying distribution of the data. As a result, the algorithm is very customizable and applicable to a wide range of problems.

iv. Support vector machines (SVMs)

Support vector machines (SVMs) and other supervised machine learning techniques can be used to solve regression and classification problems. Still, classification tasks are where SVMs are most frequently applied. SVMs operate by identifying the hyperplane in the data that most effectively divides the data points into two groups. A line or plane in n-dimensional space, where n is the number of features in the data, is called a hyperplane. SVMs locate the hyperplane where the margin between the two classes is maximized. The margin is the distance between the nearest data points in each class and the hyperplane.

v. Decision Tree

A decision tree is a supervised machine learning technique that can be used for classification as well as regression. It has a tree-like structure, with each internal node representing a dataset attribute, each branch representing a decision rule, and each leaf node representing the outcome. At each node of the tree, decision trees iteratively partition the data into subgroups based on the most significant attribute. When all of the data points in a subset belong to the same class (for classification tasks) or have the same value (for regression tasks), the splitting process ends.

vi. Gradient Boosting

One machine learning algorithm that is useful for both regression and classification problems is gradient boosting. Because the algorithm is ensemble learning, it creates a strong learner by combining several weak learners. By repeatedly training a weak learner on the residuals of the prior weak learner, gradient boosting operates. The residuals are the discrepancy between the target variable's actual values and the current weak learner's predicted values.

3.5 Implementation Requirements

The implementation requirements of liver disease risk prediction using a machine learning approach can be divided into the following steps:

1. Data collection and preparation

The first step is to collect a dataset of patients with liver disease and those without liver disease. The dataset should include relevant features such as age, sex, lab results (e.g., AST, ALT, ALB etc.), and medical history. The dataset should be cleaned and prepared by removing any outliers and missing values.

2. Feature engineering

The act of changing existing features into new features that are more informative for the machine learning model is known as feature engineering. For example, we can create new features by combining existing features (e.g., AST/ALT ratio) or by transforming continuous features into categorical features (e.g., AST level binned into high, medium, and low).

3. Model selection and training

After we have prepared the data, we can choose a machine learning model to train. There are numerous machine learning models that can be used to predict the risk of liver illness, including logistic regression, random forest, support vector machines, k-NN, decision trees, and gradient boosting. The appropriate model to use will be determined by the dataset's specific properties. Once a model has been chosen, it can be trained using the prepared data.

4. Model evaluation

Once the model has been trained, we need to evaluate its performance on a held-out test set. This will help us to assess how well the model will generalize to new data. If the model does not perform well on the test set, we may need to adjust the model parameters or try a different model.

5. Model deployment

Once we are satisfied with the performance of the model, we can deploy it to production so that it can be used to predict the risk of liver disease in new patients.

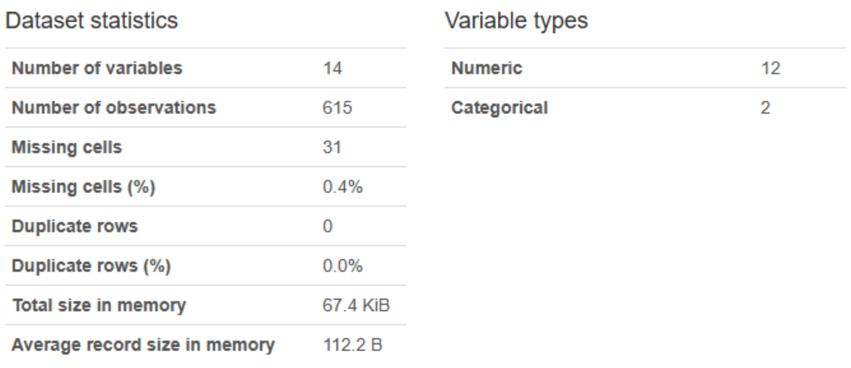
CHAPTER 4

Experimental Results and Discussion

4.1 Experimental Setup

1. Data Profiling

Statistical Analysis



Dataset statistics

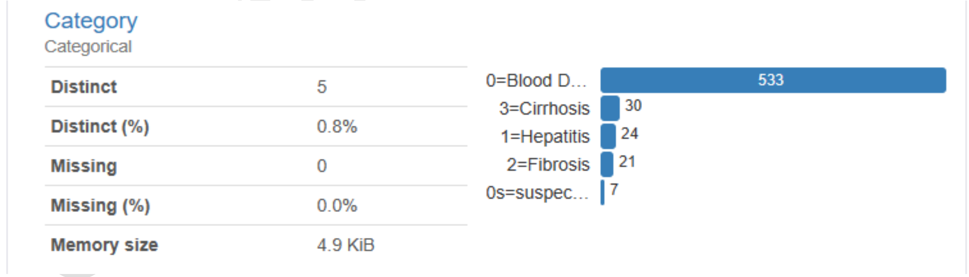
Number of variables	14
Number of observations	615
Missing cells	31
Missing cells (%)	0.4%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	67.4 KiB
Average record size in memory	112.2 B

Variable types

Numeric	12
Categorical	2

Figure 4.1 Statistical Analysis of Liver Dataset

The liver dataset has 14 variables or features. 12 features are numerical and 2 features are categorical which are ‘Category’ and ‘Sex’. There are 615 numbers of observations. There are 31 missing values and no duplicate rows.



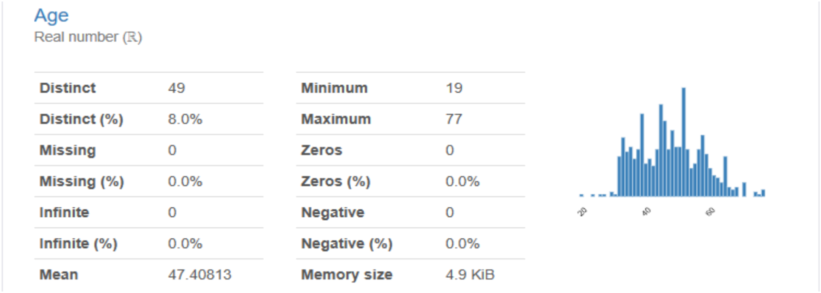
Category

Categorical

Distinct	5
Distinct (%)	0.8%
Missing	0
Missing (%)	0.0%
Memory size	4.9 KiB

Figure 4.2 Statistical Analysis of feature Category

Category feature has 5 distinct instances and has no missing values. ‘Category’ includes five categorical values that are ‘0=Blood Donor’, ‘0s=suspect Blood Donor’, ‘1=Hepatitis’, ‘2=Fibrosis’, ‘3=Cirrhosis’.

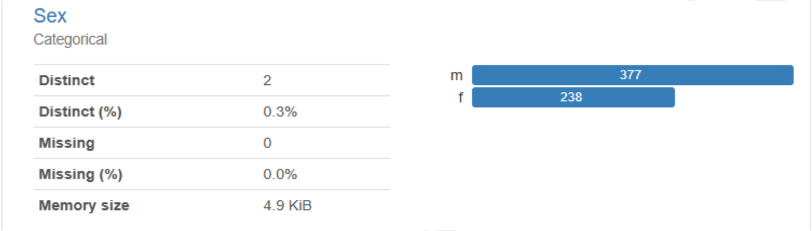


Age
Real number (ℝ)

Distinct	49	Minimum	19
Distinct (%)	8.0%	Maximum	77
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	47.40813	Memory size	4.9 KiB

Figure 4.3 Statistical Analysis of feature Age

Age feature has 49 distinct instances and has no missing value. The maximum value in the feature Age has 77 and the minimum value is 19. The mean value is 47.40813.



Sex
Categorical

Distinct	2
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	4.9 KiB

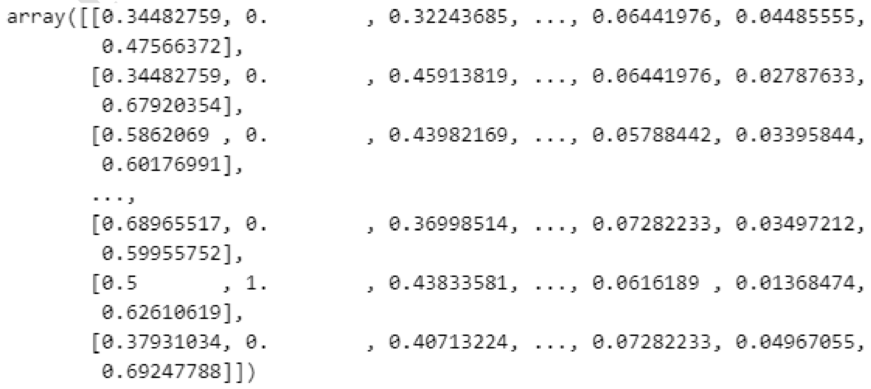
Figure 4.4 Statistical Analysis of feature Sex

Sex feature has 2 distinct instances and has no missing value. ‘Sex’ includes two categorical values that are ‘Male’ and ‘Female’. The number of male patients 377 and female patients 238.

2. Scaling

Min-Max scaling

After applying Min-Max scaling, we got a NumPy array of scaled value range of 0 to 1.

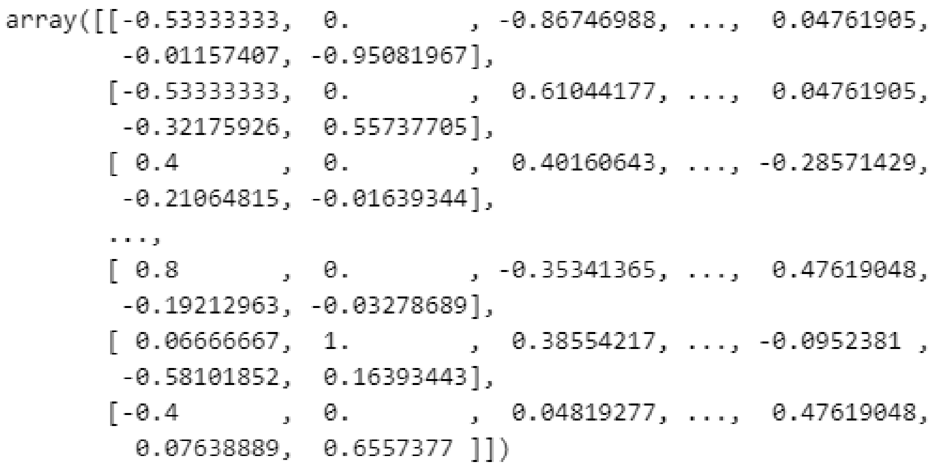


```
array([[0.34482759, 0.        , 0.32243685, ..., 0.06441976, 0.04485555,
        0.47566372],
       [0.34482759, 0.        , 0.45913819, ..., 0.06441976, 0.02787633,
        0.67920354],
       [0.5862069 , 0.        , 0.43982169, ..., 0.05788442, 0.03395844,
        0.60176991],
       ...,
       [0.68965517, 0.        , 0.36998514, ..., 0.07282233, 0.03497212,
        0.59955752],
       [0.5       , 1.        , 0.43833581, ..., 0.0616189 , 0.01368474,
        0.62610619],
       [0.37931034, 0.        , 0.40713224, ..., 0.07282233, 0.04967055,
        0.69247788]])
```

Figure 4.5 Min-Max Scaling

Robust Scaling

After applying Robust scaling, we got NumPy array of scaled value range -1 to 1

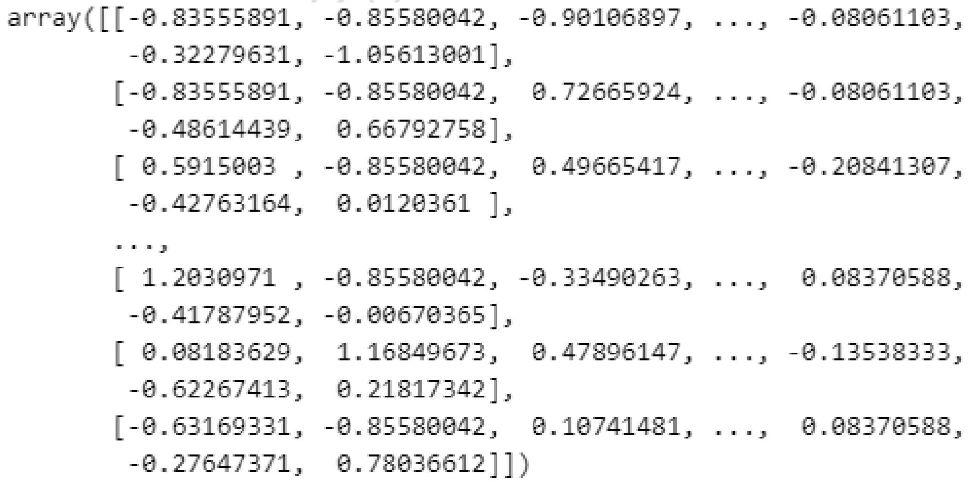


```
array([[-0.53333333,  0.        , -0.86746988, ...,  0.04761905,
        -0.01157407, -0.95081967],
       [-0.53333333,  0.        ,  0.61044177, ...,  0.04761905,
        -0.32175926,  0.55737705],
       [ 0.4       ,  0.        ,  0.40160643, ..., -0.28571429,
        -0.21064815, -0.01639344],
       ...,
       [ 0.8       ,  0.        , -0.35341365, ...,  0.47619048,
        -0.19212963, -0.03278689],
       [ 0.06666667,  1.        ,  0.38554217, ..., -0.0952381 ,
        -0.58101852,  0.16393443],
       [-0.4       ,  0.        ,  0.04819277, ...,  0.47619048,
         0.07638889,  0.6557377 ]])
```

Figure 4.6 Robust Scaling

Standard Scaling

After applying Robust scaling, we got NumPy array of scaled value range -1 to 1

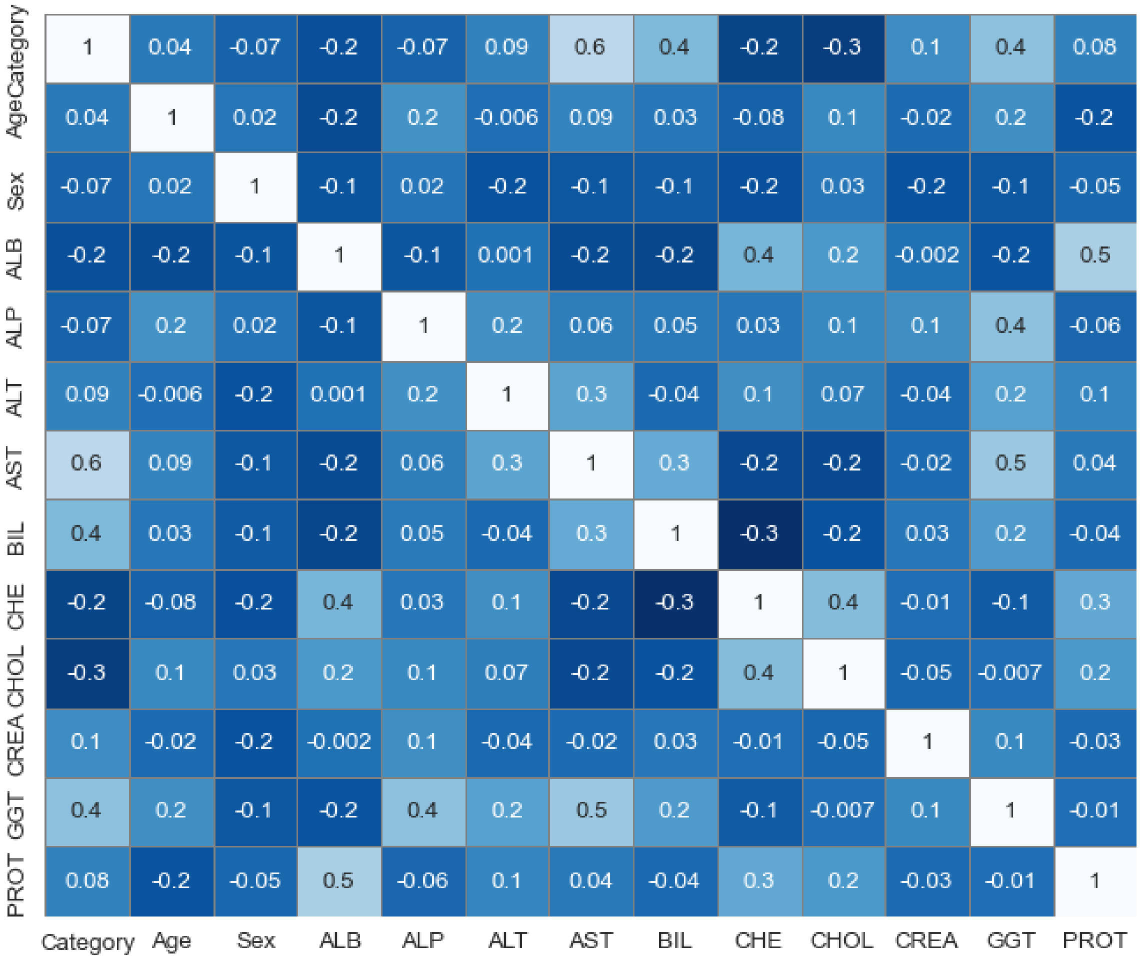


```
array([[-0.83555891, -0.85580042, -0.90106897, ..., -0.08061103,
        -0.32279631, -1.05613001],
       [-0.83555891, -0.85580042,  0.72665924, ..., -0.08061103,
        -0.48614439,  0.66792758],
       [ 0.5915003 , -0.85580042,  0.49665417, ..., -0.20841307,
        -0.42763164,  0.0120361 ],
       ...,
       [ 1.2030971 , -0.85580042, -0.33490263, ...,  0.08370588,
        -0.41787952, -0.00670365],
       [ 0.08183629,  1.16849673,  0.47896147, ..., -0.13538333,
        -0.62267413,  0.21817342],
       [-0.63169331, -0.85580042,  0.10741481, ...,  0.08370588,
        -0.27647371,  0.78036612]])
```

Figure 4.7 Standard Scaling

3. Feature Selection

The degree and direction of a linear link between two variables can be statistically measured using the correlation coefficient. The formula to compute it is the product of the standard deviations of the two variables and their covariance. A perfect negative correlation is represented by a correlation coefficient of -1, a perfect positive correlation by a correlation coefficient of 1, and no linear correlation is represented by a correlation coefficient of 0.



	Category	Age	Sex	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT
Category	1	0.04	-0.07	-0.2	-0.07	0.09	0.6	0.4	-0.2	-0.3	0.1	0.4	0.08
Age	0.04	1	0.02	-0.2	0.2	-0.006	0.09	0.03	-0.08	0.1	-0.02	0.2	-0.2
Sex	-0.07	0.02	1	-0.1	0.02	-0.2	-0.1	-0.1	-0.2	0.03	-0.2	-0.1	-0.05
ALB	-0.2	-0.2	-0.1	1	-0.1	0.001	-0.2	-0.2	0.4	0.2	-0.002	-0.2	0.5
ALP	-0.07	0.2	0.02	-0.1	1	0.2	0.06	0.05	0.03	0.1	0.1	0.4	-0.06
ALT	0.09	-0.006	-0.2	0.001	0.2	1	0.3	-0.04	0.1	0.07	-0.04	0.2	0.1
AST	0.6	0.09	-0.1	-0.2	0.06	0.3	1	0.3	-0.2	-0.2	-0.02	0.5	0.04
BIL	0.4	0.03	-0.1	-0.2	0.05	-0.04	0.3	1	-0.3	-0.2	0.03	0.2	-0.04
CHE	-0.2	-0.08	-0.2	0.4	0.03	0.1	-0.2	-0.3	1	0.4	-0.01	-0.1	0.3
CHOL	-0.3	0.1	0.03	0.2	0.1	0.07	-0.2	-0.2	0.4	1	-0.05	-0.007	0.2
CREA	0.1	-0.02	-0.2	-0.002	0.1	-0.04	-0.02	0.03	-0.01	-0.05	1	0.1	-0.03
GGT	0.4	0.2	-0.1	-0.2	0.4	0.2	0.5	0.2	-0.1	-0.007	0.1	1	-0.01
PROT	0.08	-0.2	-0.05	0.5	-0.06	0.1	0.04	-0.04	0.3	0.2	-0.03	-0.01	1

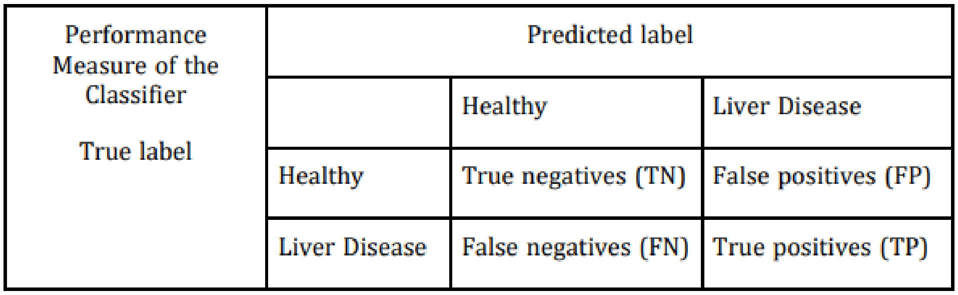
Figure 4.8 Correlation between features

4. Cross-Validation

The dataset has 615 instances. So, to tackle the problem of overfitting in machine learning we divided our dataset into two subsets. First one is train dataset and second one is test dataset. The train sub-dataset has the first 80% (492) of the dataset and the test sub-dataset has the remaining 20% (123) of the dataset.

5. Confusion Matrix

The behavioral structure of algorithms is predicted using the confusion matrix. The actual and expected class values are represented by this square matrix. The True label is represented by the columns in the confusion matrix, and the predicted label is represented by the rows. A 2*2 matrix made up of false positives (FP), false negatives (FN), true negatives (TN), and true positives (TP) is the confusion matrix in binary classification.



Performance Measure of the Classifier True label	Predicted label		
		Healthy	Liver Disease
	Healthy	True negatives (TN)	False positives (FP)
	Liver Disease	False negatives (FN)	True positives (TP)

Figure 4.9 Confusion Matrix

In this research, TP stands for the appropriately categorized cases of patients with liver disease. The value of the falsely classified cases in which the patient is not suffering from liver disease is called FP. The value of false positives for patients with liver illness is represented by FN, whereas the value of true positives for patients without liver disease is represented by TN. Six confusion matrices are produced in this study, as indicated below:

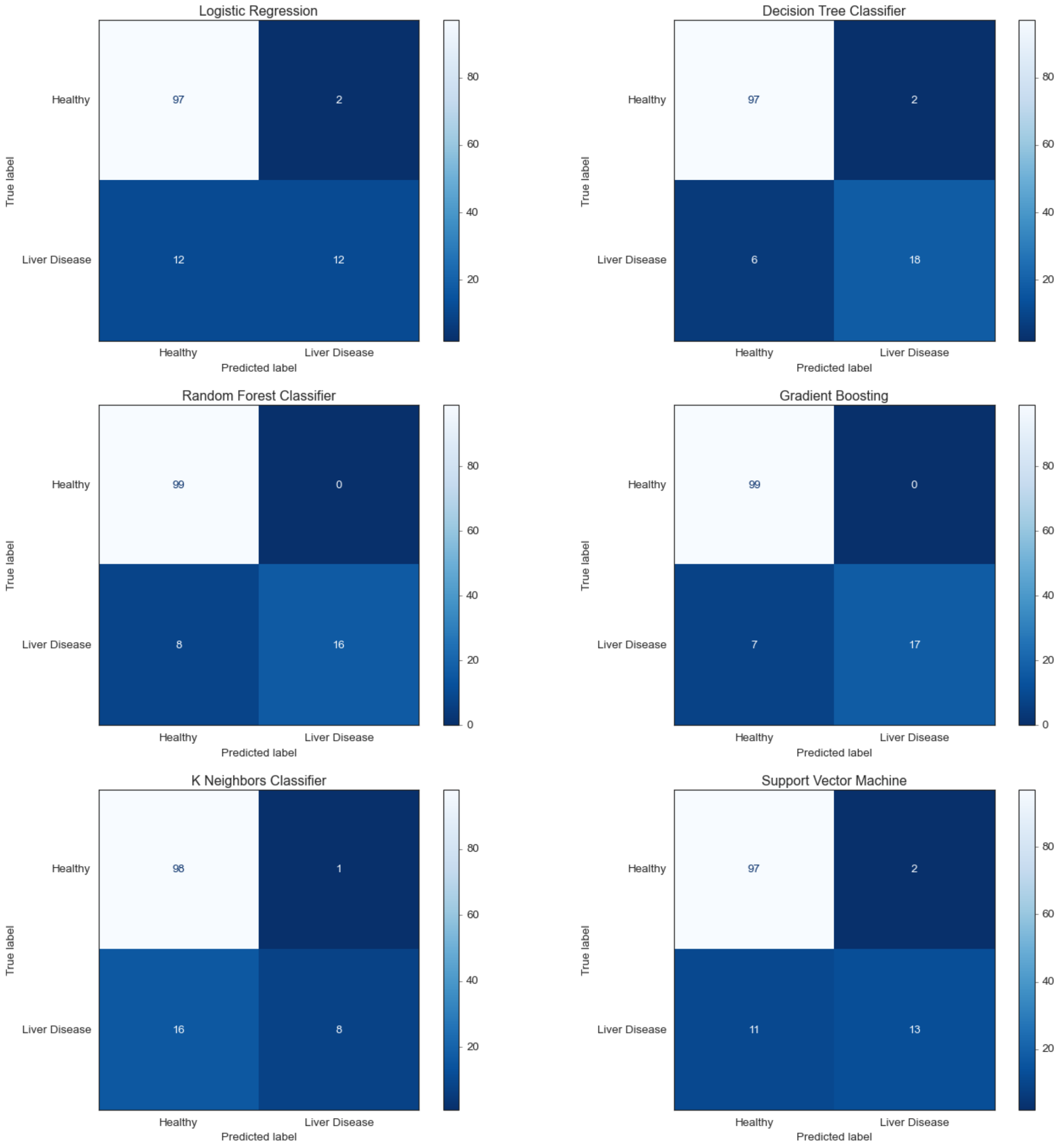


Figure 4.10 Confusion Matrices for the experimented classifiers

4.2 Experimental Results & Analysis

As described in the upward, the six classification algorithms were applied for liver disease risk prediction.

After using Min-Max scaled data (14 features) we found 94.30% accuracy by Gradient Boosting, 82.92% accuracy by Logistic Regression, 91.05% accuracy by Decision Tree, 92.68% accuracy by Random Forest, 84.55% accuracy by K Nearest Neighbors and 86.17% accuracy by Support Vector Machine on the test dataset.

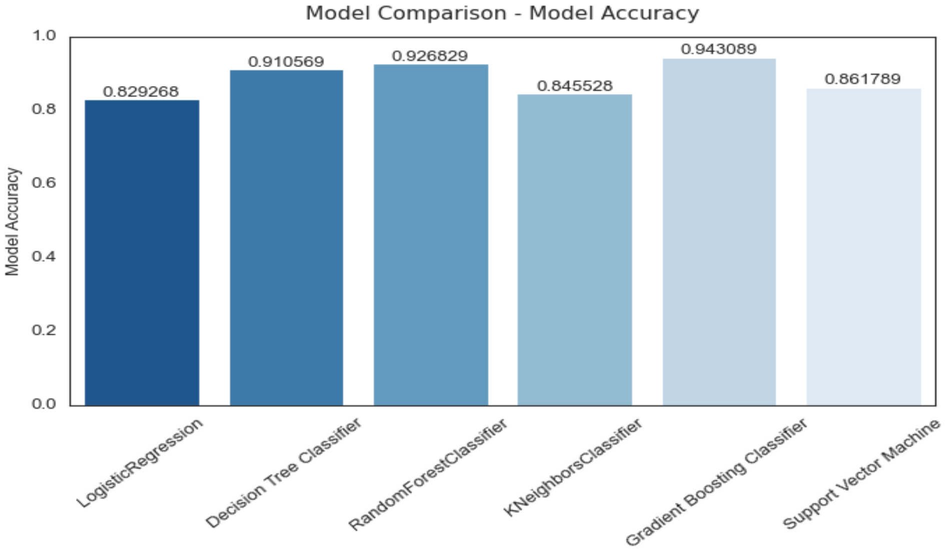


Figure 4.11 Accuracy of test set of Min-Max Scaling

After using Robust scaled data (14 features) we found 94.30% accuracy by Gradient Boosting, 88.61% accuracy by Logistic Regression, 92.68% accuracy by Decision Tree, 93.49% accuracy by Random Forest, 86.17% accuracy by K Nearest Neighbors and 89.43% accuracy by Support Vector Machine on the test dataset.

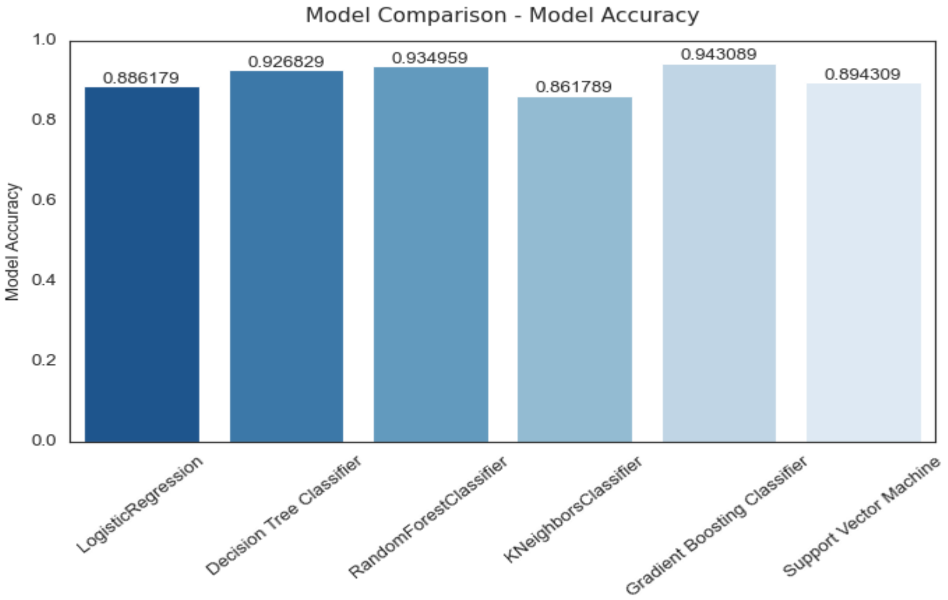


Figure 4.12 Accuracy of test set of Robust Scaling

After using Standard scaled data (14 features) we found 94.30% accuracy by Gradient Boosting, 88.61% accuracy by Logistic Regression, 92.68% accuracy by Decision Tree, 92.68% accuracy by Random Forest, 85.36% accuracy by K Nearest Neighbors and 90.24% accuracy by Support Vector Machine on the test dataset.

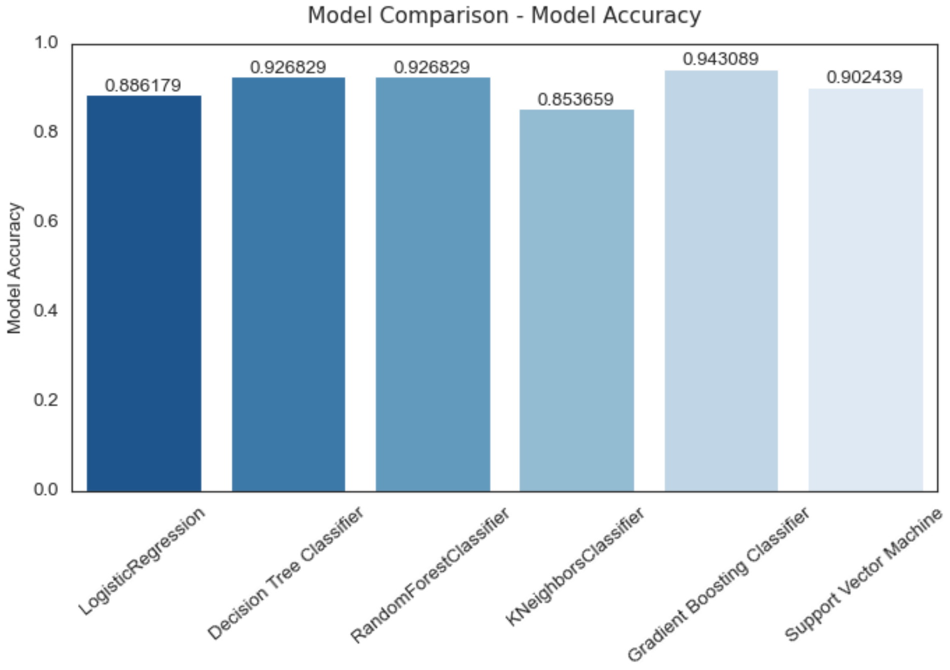


Figure 4.13 Accuracy of test set of Standard Scaling

Analyzing the performance of all classifier, it was found that in the min-max, robust and standard dataset Gradient Boosting classifier performed better as compared to all other classifier.

Table 4.2 Accuracy comparison of test set

Scaling	Gradient Boosting	Logistic Regression	Decision Tree	Random Forest	K Nearest Neighbors	Support Vector Machine
Min-Max	94.30%	82.92%	91.05%	92.68%	84.55%	86.17%
Robust	94.30%	88.61%	92.68%	93.49%	86.17%	89.43%
Standard	94.30%	88.61%	92.68%	92.68%	85.36%	90.24%

4.3 Discussion

The proper method for liver disease detection has been made available by the suggested liver disease prediction approach. Many machine learning (ML) models, such as logistic regression, support vector machines (SVMs), decision trees, random forests, k-neighbors, and gradient boosting classifier, were evaluated based on their accuracy, precision, recall, and F-measure in order to predict the occurrence of liver disease. According to the experimental findings, the gradient boosting classification method performs better than the others, having a 94% accuracy, 95% precision, and an F-measure of 94%. The others classifiers gave an acceptable score above 82%. In addition, in Table 4.3, We present the suggested models based on liver disease prediction that have been reported in research publications. In comparison to the other research, the performance of our suggested model gradient boosting is superior.

Table 4.3 Compare this study findings with other study

Study	Method	Accuracy
This study	Robust scaling, Gradient Boosting	94.3%
This study	Min-Max scaling, Gradient Boosting	94.3%
This study	Standard scaling, Gradient Boosting	94.3%
Bhupati, Deepika & Tan	Autoencoders	92.1%
Dritsas, E.; Trigka M.	Voting	80.1%
Gajendran	MAMFFN	75.3%
C. Geetha	SVM	75.04%
A.K.M Sazzdur et al.	LR	75%
Ajay & Irfan	K-NN	73.97%

CHAPTER 5

Impact on Society, Environment and Sustainability

5.1 Impact on Society

1. Improved early detection and diagnosis

Based on a person's medical history, lifestyle choices, and medical history, ML models can determine their risk of liver disease. Early detection facilitates immediate detection and treatment, preventing the disease's progression to more advanced stages.

2. Reduced healthcare costs

Early diagnosis and treatment of liver disease can significantly reduce healthcare costs by preventing the need for expensive hospitalizations, surgeries, and long-term care. ML-based risk prediction tools can help identify high-risk individuals, allowing healthcare providers to focus their resources on these patients, leading to more efficient and cost-effective care.

3. Personalized treatment plans

Healthcare professionals can design individualized treatment regimens that are specific to each patient's individual risk factors and features by using machine learning (ML) algorithms to evaluate patient data and find patterns and relationships. This personalized approach can lead to more effective treatment outcomes and improved patient satisfaction.

4. Increased public awareness

The development of ML-based liver disease risk prediction tools can raise public awareness of liver disease and its risk factors. This increased awareness can encourage individuals to adopt healthier lifestyles, seek early medical attention when needed, and participate in screening programs.

5. Empowering patients

By informing people about the possibility of liver disease, ML-based tools can empower them to take charge of their health and make informed decisions about their lifestyle and healthcare choices. This empowerment can lead to improved self-management of liver disease and overall health outcomes.

6. Reducing the burden of liver disease

Liver disease is a major global health burden, causing significant morbidity and mortality. The application of ML to predict liver disease risk has the potential to reduce this burden by promoting early detection, prevention, and effective treatment.

7. Identifying new risk factors and patterns

ML algorithms can analyze large datasets to identify new risk factors and patterns associated with liver disease. This information can be used to improve existing risk prediction models and develop new strategies for prevention and treatment.

8. Facilitating research and drug development

ML tools can be used to analyze large datasets of patient data and clinical trials, facilitating research into the causes, progression, and treatment of liver disease. This can lead to the development of new drugs and therapies for liver disease.

9. Guiding resource allocation

ML-based risk prediction tools can help healthcare systems identify high-risk populations and allocate resources more effectively, ensuring that individuals at greatest risk receive the necessary preventive care and treatment.

10. Promoting health equity

By identifying individuals at high risk of liver disease, regardless of their socioeconomic status or geographic location, ML-based tools can help promote health equity and reduce disparities in liver disease outcomes.

5.2 Impact on Environment

1. Reduced healthcare waste

Early detection and treatment of liver disease can lead to a significant reduction in healthcare waste, such as disposable medical supplies, laboratory tests, and hospital admissions. ML-based risk prediction tools can help identify high-risk individuals, allowing healthcare providers to focus their resources on these patients, leading to more efficient and environmentally friendly healthcare practices.

2. Promoting sustainable healthcare practice

By reducing the need for invasive diagnostic procedures and hospitalizations, ML-based risk prediction tools can contribute to more sustainable healthcare practices. This can lead to lower energy consumption, reduced greenhouse gas emissions, and a smaller environmental footprint for healthcare delivery.

3. Encouraging lifestyle changes

By raising awareness of liver disease risk factors, ML-based tools can encourage individuals to adopt healthier lifestyles, such as consuming a balanced diet, engaging in regular physical activity, and avoiding harmful substances. These lifestyle changes can not only improve liver health but also reduce the environmental impact of food production, transportation, and waste disposal.

4. Promoting preventive measures

ML-based risk prediction tools can help identify individuals at risk of developing liver disease due to environmental factors, such as exposure to air pollution or contaminated water. This information can be used to implement preventive measures, such as air quality monitoring, water treatment systems, and public education campaigns, to reduce environmental exposure to liver disease risk factors.

5. Supporting research into environmental causes of liver disease

ML algorithms can analyze large datasets to identify new environmental risk factors for liver disease. This information can be used to guide research into the causes and prevention of liver disease associated with environmental exposures.

6. Promoting sustainable food production practices

ML-based risk prediction tools can help identify individuals at risk of developing liver disease due to dietary factors, such as consumption of contaminated food or excessive alcohol intake. This information can be used to promote sustainable food production practices, such as reducing pesticide use, improving food safety standards, and encouraging the consumption of locally grown, organic produce.

7. Reducing the burden of liver disease on healthcare systems

The application of ML to predict liver disease risk has the potential to reduce the burden of liver disease on healthcare systems, leading to lower costs and more efficient resource allocation. This can free up resources for other environmental health initiatives.

5.3 Ethical Aspects

1. **Data privacy and security**

ML models are trained on large datasets of patient data, which includes sensitive personal information. It is crucial to ensure that this data is collected, stored, and used in a secure and ethical manner, in compliance with data privacy regulations and patient consent requirements.

2. **Fairness and bias**

ML algorithms are susceptible to biases present in the data they are trained on. These biases can lead to unfair and discriminatory outcomes, particularly for marginalized groups. It is essential to carefully evaluate ML models for bias and implement mitigation strategies to ensure fair and equitable risk prediction.

3. **Transparency and explainability**

ML models can be complex and difficult to interpret, making it challenging for healthcare providers and patients to understand the reasoning behind risk predictions. Transparency and explainability are crucial for building trust in ML-based risk prediction tools and ensuring that decisions made based on these tools are informed and accountable.

4. **Informed consent and patient autonomy**

Patients should be informed about the use of ML models in predicting their liver disease risk and should provide explicit consent for their data to be used for this purpose. Patients should also be able to access and understand their risk predictions and have the opportunity to discuss them with their healthcare providers.

5. **Clinical integration and decision-making**

ML-based risk prediction tools should be integrated into clinical practice in a way that complements and enhances the expertise of healthcare providers, rather than replacing it. Risk predictions should be considered alongside other clinical factors and patient preferences when making treatment decisions.

6. **Continuous monitoring and evaluation**

ML models should be continuously monitored and evaluated for their performance and potential biases. As new data becomes available, models should be updated and retrained to ensure their accuracy and fairness.

7. Addressing social determinants of health

ML-based risk prediction tools should not perpetuate or exacerbate social inequalities in health. Instead, they should be used to identify and address social determinants of health, such as access to healthcare, education, and healthy food options, that contribute to liver disease risk.

8. Public education and awareness

Public education and awareness campaigns should be conducted to inform individuals about the use of ML in liver disease risk prediction, the potential benefits and limitations of these tools, and the importance of informed consent and patient autonomy.

9. Regulation and oversight

Regulatory frameworks and oversight mechanisms should be developed to ensure that ML-based liver disease risk prediction tools are used ethically, responsibly, and in compliance with data privacy regulations and patient rights.

5.4 Sustainability Plan

1. Data and infrastructure sustainability

i. Data collection and storage

Implement efficient data collection methods that minimize data redundancy and ensure data quality. Utilize secure and sustainable data storage solutions that comply with data privacy regulations.

ii. Model training and deployment

Employ energy-efficient computing resources and optimize model training processes to reduce computational costs and environmental impact. Consider using cloud-based infrastructure that scales efficiently with demand.

iii. Model maintenance and updates

Establish a regular schedule for model maintenance, including retraining and evaluation, to ensure optimal performance and avoid unnecessary resource consumption.

2. Ethical and social sustainability

i. Fairness and non-discrimination

Continuously monitor and evaluate ML models for potential biases that could lead to unfair or discriminatory outcomes. Implement bias mitigation strategies and ensure that risk predictions are equitable for all individuals.

ii. Data privacy and security

Strictly adhere to data privacy regulations and implement robust cybersecurity measures to protect patient data from unauthorized access, breaches, or misuse.

iii. Transparency and explainability

Develop clear and accessible explanations of ML models and risk predictions, enabling healthcare providers and patients to understand the reasoning behind these outcomes.

iv. Informed consent and patient autonomy

Obtain explicit consent from patients before using their data for ML-based risk prediction. Provide patients with opportunities to access and understand their risk predictions and discuss them with their healthcare providers.

3. **Economic and financial sustainability**

i. **Cost-effectiveness**

Evaluate the cost-effectiveness of ML-based risk prediction tools compared to traditional methods, considering both initial development costs and ongoing maintenance expenses.

ii. **Resource allocation**

Allocate resources efficiently to ensure that ML-based risk prediction tools are used effectively and equitably across different patient populations and healthcare settings.

iii. **Return on investment (ROI)**

Demonstrate the ROI of ML-based risk prediction tools by measuring their impact on early detection, prevention, and improved patient outcomes.

iv. **Sustainable funding models**

Develop sustainable funding models for the development, maintenance, and deployment of ML-based risk prediction tools, ensuring long-term viability and accessibility.

CHAPTER 6

Summary, Conclusion, Recommendation and Implication for Future Research

6.1 Summary of the Study

Liver disease is difficult to diagnose due to the subtle nature of its symptoms, it is important to find algorithms that can more accurately predict this terrible illness. The stages in the suggested strategy for predicting the risk of liver disease offer a better alignment of each step. After the dataset has been chosen, it is balanced and the missing values are replaced as part of the preparation procedure. Following that, the accuracy using confusion matrix measurements is recorded using six different machine learning techniques (i.e., logistic regression, support vector machines (SVMs), decision trees, random forests, k-neighbors, and gradient boosting classifier). According to the experimental findings, the gradient boosting classification method performs better than the others, having a 94% accuracy, 95% precision, and an F-measure of 94%. The goal of this work was to use machine learning techniques to predict the risk of liver disease early. The findings of this study could help important decision-makers and medical experts identify liver disease risks early.

6.2 Conclusions

Liver disease is a major global health concern that affects millions of people globally. Conventional diagnostic techniques, which are frequently intrusive and expensive, impede early diagnosis and prompt action, which has a negative impact on patient outcomes. With the ability to identify those at high risk early on, machine learning (ML) has emerged as a promising method for liver disease risk prediction. This could lead to protective treatments and better patient outcomes. Machine Learning (ML) algorithms can effectively analyze vast amounts of medical data, extracting patterns and relationships that can predict the likelihood of developing liver disease. This study has demonstrated the efficacy of ML in predicting various liver disease types, including hepatitis c, liver cirrhosis and fibrosis. ML algorithms can achieve high prediction accuracy, with some studies exceeding 90%, outperforming traditional methods.

6.3 Implication for Further Study

1. Developing standardized data collection methods

To ensure consistent and reliable results, it is crucial to establish standardized data collection methods across studies. This will involve defining common data elements, ensuring data quality control, and harmonizing data formats. Standardized data collection will facilitate comparisons between studies and enhance the generalizability of ML models.

2. Enhancing interpretability of ML models

Despite their great prediction accuracy, machine learning models may not be widely accepted in clinical practice due to their interpretability issues. Further research should focus on developing methods to improve the interpretability of ML models, making it easier for healthcare professionals to understand and trust their predictions. This could involve techniques such as explainable AI (XAI) and feature importance analysis.

3. Validating ML models in real-world clinical settings

Evaluating the performance of ML models in real-world clinical settings is essential to assess their generalizability and effectiveness in practice. This involves conducting prospective studies in diverse patient populations, considering factors such as healthcare provider expertise, patient adherence, and real-world data limitations.

4. Explore specific disease progression and treatment outcomes

The potential of ML in predicting specific liver disease progression stages and treatment outcomes holds great promise. ML models that can identify people at risk of rapid progression or less than ideal treatment outcomes should be the main focus of research. This data may help generate individualized treatment plans and enhance patient results.

References

[1] Dritsas, E.; Trigka, M. Supervised Machine Learning Models for Liver Disease Risk Prediction. Computers 2023, 12, 19. https://doi.org/10.3390/computers12010019

[2] Rakshith D B , Mrigank Srivastava , Ashwani Kumar, Gururaj S P, 2021, Liver Disease Prediction System using Machine Learning Techniques, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 10, Issue 06 (June 2021)

[3] Bhupathi, Deepika & Tan, Christine Nya-Ling & Sremath Tirumala, Sreenivas & Ray, Sayan. (2022). Liver disease detection using machine learning techniques.

[4] Gajendran, G.; Varadharajan, R. Classification of Indian liver patient's data set using MAMFFN. In Proceedings of the AIP Conference Proceedings, Coimbatore, India, 17–18 July 2020; Volume 2277, p. 120001.

[5] C. Geetha and A. Arunachalam, "Evaluation based Approaches for Liver Disease Prediction using Machine Learning Algorithms," 2021 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2021, pp. 1-4, doi: 10.1109/ICCCI50826.2021.9402463.

[6] Rahman, A.K.M Sazzadur et al. "A Comparative Study On Liver Disease Prediction Using Supervised Machine Learning Algorithms." International Journal of Scientific & Technology Research 8 (2019): 419-422.

[7] Kannapiran, Thirunavukkarasu & Singh, Ajay & Irfan, Md & Chowdhury, Abhishek. (2018). Prediction of Liver Disease using Classification Algorithms. 1-3. 10.1109/CCAA.2018.8777655.

[8] Arias, I.M.; Alter, H.J.; Boyer, J.L.; Cohen, D.E.; Shafritz, D.A.; Thorgeirsson, S.S.; Wolkoff, A.W. The Liver: Biology and Pathobiology; John Wiley & Sons: Hoboken, NJ, USA, 2020.

[9] Rycroft, J.A.; Mullender, C.M.; Hopkins, M.; Cutino-Moguel, T. Improving the accuracy of clinical interpretation of serological testing

[10] Lambrecht J and Tacke F (2021) Controversies and Opportunities in the Use of Inflammatory Markers for Diagnosis or Risk Prediction in Fatty Liver Disease. Front. Immunol. 11:634409. doi: 10.3389/fimmu.2020.634409

[11] Fazakis, N.; Kocsis, O.; Dritsas, E.; Alexiou, S.; Fakotakis, N.; Moustakas, K. Machine learning tools for long-term type 2 diabetes risk prediction. IEEE Access 2021, 9, 103737–103757.

[12] Dhamodharan, S. Liver Disease Prediction Using Bayesian Classification. 2016. Available online: https://www.ijact.in/index. php/ijact/article/viewFile/443/378 (accessed on 02 November 2023).

[13] Srivastava, A.; Kumar, V.V.; Mahesh, T.; Vivek, V. Automated Prediction of Liver Disease using Machine Learning (ML) Algorithms. In Proceedings of the 2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Bhilai, India, 21–22 April 2022; pp. 1–4.

[14] Choudhary, R.; Gopalakrishnan, T.; Ruby, D.; Gayathri, A.; Murthy, V.S.; Shekhar, R. An Efficient Model for Predicting Liver Disease Using Machine Learning. In Data Analytics in Bioinformatics: A Machine Learning Perspective; Wiley Online Library: Hoboken, NJ, USA, 2021; pp. 443–457.

[15] Lichtinghagen,Ralf, Klawonn,Frank, and Hoffmann,Georg. (2020). HCV data. UCI Machine Learning Repository. https://doi.org/10.24432/C5D612

[16] Kumar, P.; Thakur, R.S. Early detection of the liver disorder from imbalance liver function test datasets. Int. J. Innov. Technol. Explor. Eng. 2019, 8, 179–186.

[17] Idris, K.; Bhoite, S. Applications of machine learning for prediction of liver disease. Int. J. Comput. Appl. Technol. Res 2019, 8, 394–396. Computers 2023, 12, 19 15 of 15

[18] Muthuselvan, S.; Rajapraksh, S.; Somasundaram, K.; Karthik, K. Classification of liver patient dataset using machine learning algorithms. Int. J. Eng. Technol. 2018, 7, 323

[19] Center for Liver Disease and Transplantation, available at Liver Functions, Location, Anatomy and Disease | Columbia Surgery, last accessed on 27 November 2023 at 12.22 pm

[20] Modifiable & Non-Modifiable Risk Factors for Heart Disease, available at Modifiable & Non-Modifiable Risk Factors for Heart Disease | Amy Myers MD, last accessed on 27 November 2023 at 01.13 pm

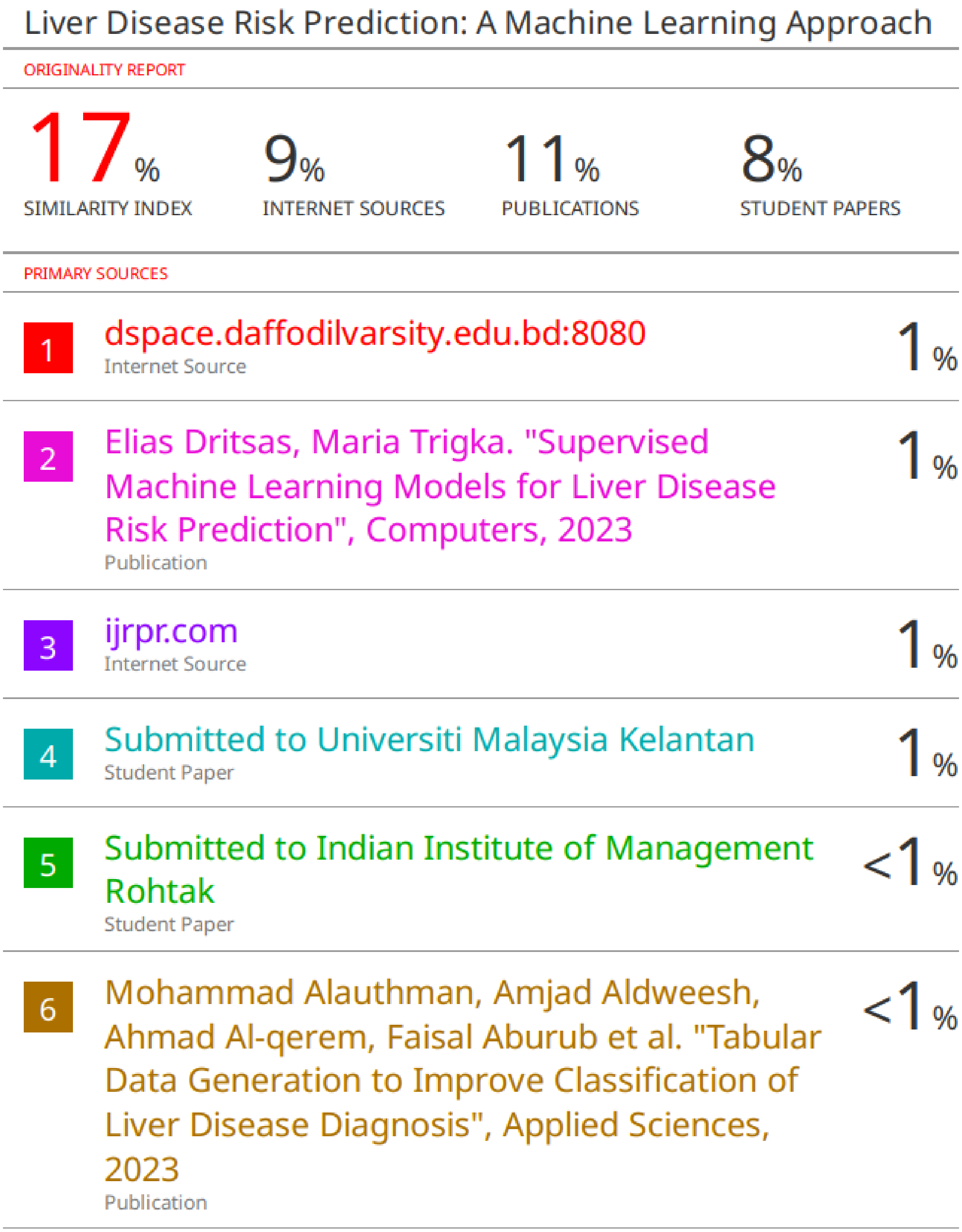
[21] Prediction Model of Adverse Effects on Liver Functions of COVID-19 ICU Patients, available at Prediction Model of Adverse Effects on Liver Functions of COVID-19 ICU Patients - PMC (nih.gov) , last accessed on 27 November 2023 at 02.33 pm

[22] FEATURE SELECTION TECHNIQUES IN MACHINE LEARNING [2023 EDITION] , available at Feature Selection Techniques in Machine Learning [2023 Edition] - Dataaspirant, last accessed on 27 November 2023 at 06.20 pm

[23] Feature Selection Techniques in Machine Learning, available at Feature Selection Techniques in Machine Learning - GeeksforGeeks, last accessed on 27 November 2023 at 06.30 pm

[24] k-Nearest Neighbors(k-NN) in Machine Learning, available at k-Nearest Neighbors(k-NN) in Machine Learning - Spark By {Examples} (sparkbyexamples.com) , last accessed on 27 November 2023 at 06.52 pm

[25] Bahramirad, S.; Mustapha, A.; Eshraghi, M. Classification of liver disease diagnosis: A comparative study. In Proceedings of the 2013 Second International Conference on Informatics & Applications (ICIA), Lodz, Poland, 23–25 September 2013; pp. 42–46.



Liver Disease Risk Prediction: A Machine Learning Approach

ORIGINALITY REPORT

17%	9%	11%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
2	Elias Dritsas, Maria Trigka. "Supervised Machine Learning Models for Liver Disease Risk Prediction", Computers, 2023 Publication	1%
3	ijrpr.com Internet Source	1%
4	Submitted to Universiti Malaysia Kelantan Student Paper	1%
5	Submitted to Indian Institute of Management Rohtak Student Paper	<1%
6	Mohammad Alauthman, Amjad Aldweesh, Ahmad Al-qerem, Faisal Aburub et al. "Tabular Data Generation to Improve Classification of Liver Disease Diagnosis", Applied Sciences, 2023 Publication	<1%