Project Report On

LIVER DISEASE PROGRESSION USING MACHINE LEARNING

Submitted in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in

APPLIED STATISTICS AND DATA ANALYTICS

by

NAVYA N

(Register No. SM23AS010)

(2023-2025)

Under the Supervision of

SMT. VISMAYA VINCENT

DEPARTMENT OF MATHEMATICS
ST. TERESA'S COLLEGE (AUTONOMOUS)
ERNAKULAM, KOCHI - 682011
APRIL 25

ST TERESA'S COLLEGE (AUTONOMOUS), ERNAKULAM

THE FEAR OF THE LORD IS THE BEGINNING OF WISDOM

ST. TERESA'S COLLEGE
(AUTONOMOUS)
ESTD 1925

CERTIFICATE

This is to certify that the dissertation entitled, **LIVER DISEASE PROGRESSION USING MACHINE LEARNING** is a bonafide record of the work done by **NAVYA N** under my guidance as partial fulfilment of the award of the degree of **Master of Science in Applied Statistics and Data Analytics** at St. Teresa's College (Autonomous), Ernakulam affiliated to Mahatma Gandhi University, Kottayam. No part of this work has been submitted for any other degree elsewhere.

Date:

Place: Ernakulam

Vismaya

VISMAYA VINCENT
Assistant Professor,
Department of Mathematics and Statistics
St. Teresa's College (Autonomous),
Ernakulam

Nisha Oommen
Assistant Professor & HOD
Department of Mathematics and Statistics (SF)
St. Teresa's College (Autonomous)
Ernakulam

ST.TERESA'S COLLEGE (AUTONOMOUS)
DEPARTMENT OF MATHEMATICS AND STATISTICS
★ ERNAKULAM ★

External Examiners

1. Sangeetha Chandran
30.04.2025

2. Anju N.B
30 04 25

DECLARATION

I hereby declare that the work presented in this project is based on the original work done by me under the guidance of **SMT. VISMAYA VINCENT**, Assistant Professor, Department of Mathematics and Statistics, St. Teresa's College (Autonomous), Ernakulam and has not been included in any other project submitted previously for the award of any degree.

NAVYA N

Ernakulam

SM23AS010

Date: 30|4|25

ACKNOWLEDGEMENT

I take this opportunity to thank everyone who has encouraged and supported me to carry out this project.

I am very grateful to my project guide Ms. Vismaya Vincent for her immense help during the period of work.

In addition, I acknowledge with thanks to the department for all the valuable support and guidance that has significantly contributed to the successful completion of this project.

I would also like to thank the HOD for her valuable suggestions and critical examinations of the project.

NAVYA N

SM23AS010

Place: Ernakulam

Date: 30|04|25

ABSTRACT

Liver Disease is a major worldwide health concern, liver disease requires precise predictive models for both early diagnosis and progression analysis. This work uses the Liver Disease Patient data-set to examine the course of liver disease using machine learning and deep learning techniques. K-Means Clustering was used to identify patterns among patients after Principal Component Analysis (PCA) was used to reduce dimension. Support Vector Machines (SVM) and Logistic Regression were used to assess the models. Additionally, Random Forest and Gradient Boosting were used to identify the key characteristics of liver illness. Deep learning models, such as a Wide and Deep Neural Network (WDNN) and a Fully Connected Neural Network (FCNN), were used to improve predictive performance. To determine the best method for predictive analysis, the models were compared. The significance of combining deep learning and machine learning for reliable liver disease prediction is highlighted by this work, which will help with early identification and better patient outcomes.

ST.TERESA'S COLLEGE (AUTONOMOUS)
ERNAKULAM

Certificate of Plagiarism Check for Dissertation

Author Name	NAVYA N
Course of Study	MSc. Applied Statistics & Data Analytics
Name of Guide	Ms. Vismaya Vincent
Department	PG. Dept. of Mathematics & Statistics
Acceptable Maximum Limit	20
Submitted By	library@teresas.ac.in
Paper Title	LIVER DISEASE PROGRESSION USING MACHINE LEARNING
Similarity	9% AI-9%
Paper ID	3420925
Total Pages	49
Submission Date	2025-03-21 10:36:01

Signature of Student

Signature of Guide

Anu Maria

Checked By
College Librarian

* This report has been generated by DrillBit Anti-Plagiarism Software

CONTENTS

Page no.

Chapter 1

Introduction

Liver disease is a significant public health concern, Due to the high incidence of cirrhosis, hepatitis, and liver cancer, liver disease is a serious public health concern. Timely interventions, better patient outcomes, and lower healthcare costs all depend on accurate liver disease progression prediction. With one out of five persons affected, liver disease is rapidly spreading like an epidemic. India has seen an alarming rise in liver-related fatalities, with 268,580 deaths (3.17% of all deaths) annually, accounting for 18.3% of the 2 million liver-related deaths worldwide. It has been demonstrated that machine learning algorithms can accurately forecast the course of diseases. Predicting liver disease by the analysis of patient data, such as clinical factors (blood tests, demographics), and medical imaging is becoming more and more common with Machine Learning (ML) and Deep Learning (DL) algorithms.

Studies that attempted to address this gap have been conducted. Health care professionals' ability to raise patient awareness and intervene promptly is greatly impacted by this. This foretells the development of liver disease. Abdalrada (2019) states that the logistic regression method is the foundation of the suggested prediction model. liver disease prediction machine learning models, such as ensemble classifiers (XG Boost, Bagging, and RF) and single classifiers (SVM and GP). The liver disorder data set's feature importance is also examined. In the performance evaluation by Zhao (2022) presents F1-score, recall, precision, accuracy, and balanced.
It has been discovered that a lower A/G level is linked to a higher risk of SAP and liver disease. The results emphasized the need to implement more suitable methods to reduce the prevalence of SAP in patients with lower A/G and the crucial role that A/G plays in predicting SAP in patients with AIS. It will be crucial for future studies to examine how nutritional interventions that increase the A/G ratio affect the results of AIS and SAP (Chen, 2022). the effectiveness of machine learning algorithms in using the Liver Patient Dataset to forecast the risk of liver disease. This could assist patients in choosing prompt and efficient care (Telaq, 2021).

Through a comparison of the effectiveness of machine learning (ML) and deep learning (DL) algorithms, the suggested work seeks to lower the high cost of liver disease diagnosis through detection. Several methods, such as Convolution Neural Network (CNN), Artificial Neural Network (ANN), Gaussian Naive Bayes (GNB), Random Forest (RF), and Logistic Regression (LR), have been used in the suggested study. Accuracy, precision, recall, F-1 score, train time, and test time are among the measures used to evaluate the performance's efficacy. The main focus of the proposed work is on using medical data to identify liver-related diseases. (Bhusnurmath, 2024) In vertebrates, the liver is a vital organ for metabolism, performing vital functions like detoxification, protein synthesis, and assistance in

The proposed work aims to reduce the high cost of liver disease diagnosis through detection by comparing the efficiency of machine learning (ML) and deep learning (DL) algorithms. In the proposed study, variety of algorithms have been employed that include Convolution Neural Network (CNN), Artificial Neural Network (ANN), Gaussian Naive Bayes (GNB), Random Forest (RF), and Logistic Regression (LR). The effectiveness of the performance is assessed using a variety of metrics that include accuracy, precision, recall, F-1 score, train time, and test time. Proposed work primarily focuses on the use of medical data for the detection of disease related to liver. (Bhusnurmath, 2024) The liver is an essential organ for metabolism in vertebrates, carrying out important tasks such as detoxifying, making proteins, and aiding in digestion. It controls the metabolism of carbohydrates, produces hormones, stores calories, and breaks down red blood cells. It is located in the upper right section of the abdomen, beneath the diaphragm. It produces bile, a crucial component of digestion, which aids in the breakdown of fat. Bile is collected and released into the duodenum by the gallbladder, which is situated underneath the liver's right lobe. Hepatocytes, the primary cells of the liver, carry out a variety of biochemical tasks, producing and decomposing essential chemicals required for bodily functioning. It controls how nutrients are processed, eliminates toxins, and boosts immunity. The liver, often known as the body's chemical factory, performs about 500 vital functions and is crucial for maintaining equilibrium and health.

A comprehensive approach involving early diagnosis, lifestyle modifications, and medical interventions is essential to managing and preventing the progression of liver disease.

1.1Liver Diseases

Numerous factors, such as infections, medicines, alcohol, or genetics, might contribute to liver disease. Typical forms of liver illness include the following: Abnormal bilirubin, albumin, and enzyme levels are signs of liver cirrhosis.

1.Liver Cirrhosis

Cirrhosis of the liver is a situation in which healthy liver cells turn into scar tissue, resulting in a decline in liver function. This condition may occur due to longterm alcohol consumption, viral infections like hepatitis, or a buildup of fat in the liver. Signs of cirrhosis are often reflected in abnormal levels of bilirubin, albumin, and liver enzymes.

2.Hepatitis (A, B, C, etc.)

Liver inflammation, frequently brought on by viral infections, is referred to as hepatitis. It can manifest as either a short-term or long-term sickness, and it can eventually cause damage to the liver. The key to diagnosing hepatitis is having high levels of bilirubin and liver enzymes. The various forms—A, B, and C—have different modes of transmission and differing levels of severity.

3.Fatty Liver Disease

When the liver accumulates too much fat, fatty liver disease results. It is frequently linked to diseases like obesity, diabetes, or poor eating habits and can be brought on by alcohol (AFLD) or occur without alcohol usage (NAFLD). If left untreated, it could cause inflammation and liver damage.

4.Liver Cancer (Hepatocellular Carcinoma)

Hepatocellular carcinoma, a form of liver cancer, typically arises in people with cirrhosis or long-term hepatitis. It is linked to significant changes in enzyme levels and is frequently identified at later stages. Recognizing it early can enhance treatment results.

5.Cholestasis (Bile Flow Obstruction)

Cholestasis happens when bile does not flow properly from the liver, which leads to a buildup of bile within the organ. This condition can be triggered by gallstones, liver ailments, or the side effects of medications. Elevated alkaline phosphatase levels along with jaundice are important signs.

6.Liver Failure

A serious disorder known as liver failure occurs when the liver is unable to function normally. Infections, dangerous drugs, long-term diseases, or unexpected liver injury can all cause this. Numerous abnormal liver function tests are indicative of the onset of the illness and necessitate prompt medical intervention.

1.2 Objectives of the study

1. To Investigate the effectiveness of various machine learning methods, including PCA for dimensionality reduction, K-means clustering for pattern detection, and Logistic Regression and SVM for classification.

2.To Utilize Gradient Boosting and Random Forest models to predict liver disease while identifying the most important features influencing the diagnosis.

3.To Evaluate the performance of deep learning models like Fully Connected Neural Networks (FCNNs) and Wide & Deep Neural Networks (WDNNs) in diagnosing liver disease.

Chapter 2
Literature Review

Analysis from the previous research paper related to this paper uses various clinical datasets and makes predictions using advanced statistical techniques, AI technologies, machine learning algorithms, deep learning techniques, etc.

Abdalrada et al. (2019) In this study, suggest using a logistic regression model to calculate the likelihood of developing the disease with the help of the ILPD dataset. The model provided an accuracy rate of 72. 4%, sensitivity of 90. 3%, and specificity of 78. 3%. Important factors that influence the predictions are Age, Direct Bilirubin, SGPT, Total Proteins, and Albumin. This method aids both physicians and patients in effectively tracking the progression of the illness.

Moustafa et al. (2021) The purpose of this study was to use machine learning algorithms on medical data from 615 people in order to find significant predictors of liver disease. Principal Component Analysis (PCA) decreased dimensionality, while Multiple Imputations by Chained Equations (MICE) addressed missing values found by data visualizations. Important factors were ordered by the Gini index, which validated important predictors. Artificial Neural Networks, Random Forest (RF), and Support Vector Machines were used to classify the dataset, which was divided into training (399 samples) and testing (216 samples). Class imbalance was resolved by the Synthetic Minority Oversampling Technique (SMOTE). With an accuracy of 98.14% (p<0.001), Random Forest outperformed the other models. These findings show how well machine learning (ML) may predict liver illness and enhance inference-based diagnosis.

Mutlu et al. (2021) A serious worldwide health issue, liver disease necessitates precise predictive models for early detection. In order to improve classifier performance for liver disease diagnosis, this study investigates Bayesian optimization. It assesses Random Forest, SVM, AdaBoost, and XG Boost using a Kaggle dataset from the UCI library, using Pearson Correlation Feature Selection to find the best features. By adjusting hyperparameters, Bayesian optimization increases model accuracy. According to the results, RF had the highest accuracy (81.06%), followed by SVM (80.81%), XG Boost (79.85%), and AdaBoost (77.08%). This indicates how well optimization works to improve the prediction of liver illness.

Zhao et al. (2022) In order to reduce data randomness, leave-one-out cross-validation is used in this paper's machine learning models for liver disease prediction. It examines the significance of features and their interrelationships. With an accuracy of 80.35%, Random

Forest beats other models in the experiment, most likely as a result of its ensemble approach, which lessens overfitting from unbalanced data. According to the study, Random Forest is a useful tool for helping medical professionals diagnose liver illness.

Chen et al. (2022) The study revealed that a low A/G level was associated with an increased SAP risk. Appropriate preventative measures for SAP should be taken in AIS patients with a low A/G level. The logistic regression model was used to determine the association between A/G and SAP, and a forest plot was drawn. Compared with the non-SAP group, the SAP group had a lower A/G level ($P < 0.001$). Then, A/G was divided into quartiles. In comparison to Q3 (A/G = 1.25–1.39), logistic regression revealed that patients with a lower A/G (A/G $\leq$ 1.09) had a higher risk of SAP (OR = 1.96, 95% CI, 1.56–2.46, $P < 0.001$). On the contrary, those with a higher A/G (A/G $\geq$ 1.4) had a lower SAP risk (OR = 0.73, 95% CI, 0.54–0.97, $P =$ 0.029).

Gupta et al. (2022) The study predicts liver illness using the Liver Patient dataset from the UCI Repository (Supervised Learning). Algorithms such as Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbours, Gradient Boosting, Extreme Gradient Boosting, and Light GBM are used to analyse medical data from hospital patients. Following feature selection, the results demonstrate that these models had good accuracy, indicating their potential to enhance the prediction of liver disease.

Joshua et al. (2023) Liver disease is on the rise worldwide, but it frequently goes undiagnosed until it is too late because of its limited symptoms. Accurate prediction is crucial since early discovery can save lives. Using the Indian Liver Patient Dataset (ILPD), this work suggests an ensemble learning-based model with improved preprocessing for liver disease prediction. Data balance, feature scaling, and selection strategies were used in the application of six ensemble learning algorithms: Gradient Boosting, XG Boost, Bagging, Random Forest, Extra Trees, and Stacking. Multivariate imputation was used to manage missing values, and log1p and other scaling techniques were used to modify skewed data. The efficiency of the suggested method for detecting liver disease was demonstrated by the Extra Trees and Random Forest classifiers, which obtained the maximum testing accuracy of 91.82% and 86.06%, respectively.

Ganie et al. (2024) Millions of people worldwide suffer from chronic liver disease (CLD), hence early detection is essential for prompt treatment. Preventive healthcare can be supported by machine learning, which can also improve diagnosis and results. However, ensemble learning can overcome the drawbacks of classical machine learning, providing predictions that are more dependable and accurate.

Sompura et al. (2024) An essential component of automated disease detection is the extraction of valuable information from massive medical datasets. Liver disease has become one of the most prevalent illnesses worldwide in recent years. This research proposes a model for liver disease diagnosis based on Convolutional Neural Networks (CNNs). Additionally, CNN's performance was contrasted with that of more conventional machine learning techniques, such as Logistic Regression (LR), K-nearest Neighbours (KNN), Support Vector Machine (SVM), and Naive Bayes (NB). BUPA and ILPD were the two datasets used for evaluation. According to the experimental results, CNN was successful in classifying liver disease, achieving 75.55% and 72.00% accuracy on the BUPA and ILPD datasets, respectively.

Betageri et al. (2024) Alcoholism, obesity, and untreated hepatitis are some of the leading causes of liver disease, a serious worldwide health concern with a complicated and expensive diagnosis. By contrasting machine learning (ML) with deep learning (DL) models for liver disease diagnosis, this study seeks to lower diagnostic costs. Accuracy, precision, recall, F1-score, training time, and testing time are used to assess performance, and CNN, ANN, GNB, RF, and LR are used. The study shows that the suggested models perform better than current methods using medical data. The results demonstrate how well ML and DL work to enhance early diagnosis and detection, improve liver disease prediction, and make it more effective, accessible, and economical—all of which contribute to prompt medical intervention and improved patient outcomes.

Chapter 3
Materials and Methodology

3.1Dataset

This study made use of the Liver Disease Patient dataset, sourced from open-access website Kaggle, consists of 30,691 patient records collected from Electronic Health Records (EHR), making it a valuable resource for medical research and predictive modelling. The dataset includes demographic details such as age and gender, along with clinical biomarkers like total and direct bilirubin, alkaline phosphate, aspartate aminotransferase (SGOT), alanine aminotransferase (SGPT), albumin, and the albumin-to-globulin ratio. The dataset divides into patients with liver disease (21,917cases) and patients without liver disease (8,774 cases) uses 1 for "Disease Present" and 2 for "No Disease". These features are critical for evaluating liver function and identifying potential liver diseases.

3.2 Machine Learning

Machine learning, a branch of computer science and artificial intelligence (AI), specializes in modelling human learning processes using data and algorithms, subsequently gaining accuracy. Because of his studies on the game of checkers, Arthur Samuel is credited with coining the term "machine learning." In 1962, Robert Nealey, who considered himself a master of checkers, faced out against IBM's 7094 computers and lost. Thus, in light of current capabilities, his accomplishments fade in comparison, but they are still regarded as key moments in the history of artificial intelligence. Advancements in storage and processing capacities have led to the development of machine learning-based products, including Netflix recommendation algorithms and self-driving automobiles.

Machine learning is one of the key elements in the rapidly expanding disciplines of data science. For gaining import insights in data mining projects, algorithms are trained using statistical techniques for building the classification and predictive model. The decisions made as a result of these insights serve as a key growth indicator in enterprises and applications. Data Scientists are more in demand as big data is developing and flourishing.

They were expected to assist in determining business questions and with information needed to address them. The accelerating solution development frameworks like TensorFlow and Py Torch are utilized for machine learning algorithm development.

Types of Machine Learning

3.2.1 Supervised Learning

In machine learning, supervised learning was developed using labelled datasets and based upon supervision. In other words, every piece of input data has a matching goal or label. Learning a mapping from inputs to outputs that provide the models is the aim of supervised learning to accurately categorize or predict the fresh, unknown dataset. One of the main characteristics of supervised learning is that it requires a labelled dataset as input and produces labelled output. Reducing the discrepancy between the expected and real labels is the ultimate goal of this kind. Among the benefits of supervised machine learning are its versatility, ability to achieve high accuracy with a sufficient amount of labelled data, and applications across other disciplines.

Supervised machine learning is commonly used in image classification to identify photos and identify items in the images, emotional analysis to forecast demographics like population growth or health indicators using regression and ascertain the sentiment conveyed in a text, in the diagnosis of illnesses using patient information, in language translation for the purpose of translating text between languages, in medicine, etc. Algorithms that can be applied to supervised learning include decision trees, neural networks, support vector machines, and linear regression. The benefits of supervised machine learning include versatility, applications in a variety of fields, and high accuracy with adequate labelled data.

3.2.2 Unsupervised Learning

Training the model on an unlabelled dataset (one without target labels) is known as unsupervised learning. Since it was constructed without any oversight, the algorithm is given an input dataset that has been trained to group items based on shared qualities rather than rewarding or optimizing to a certain result. Unlabelled, unclassified data is used to train models, which then act independently on the data. The goal is to classify or group the unsorted dataset in order to uncover hidden patterns, structures, and relationships. This is helpful for exploratory data analysis and uncovering hidden patterns.

A clustering strategy groups data points based on similarities, while a dimensionality reduction technique reduces the number of characteristics while maintaining key features. Principal component analysis (PCA), autoencoders, K-means clustering, hierarchical clustering, and others are a few examples of methods used in unsupervised learning. Customer segmentation, which groups clients according to their purchase behaviours, topic modelling, anomaly detection, which finds odd patterns in data, and the extraction of underlying themes from document collections are a few uses for this technique.

3.2.3 Semi-supervised Learning

Supervised and unsupervised learning are combined to create semi-supervised learning. For model training, it contains a small amount of labelled data and a big amount of unlabelled data. Since both supervised and unsupervised learning rely on the existence or absence of labels, this kind differed from others. The shortcomings of both supervised and unsupervised learning approaches are intended to be addressed by this. Using all available data, the dataset is initially grouped using an unsupervised learning methodology, after which it can be classified based on the labelled data. Even when there is a shortage of labelled data, using the unlabelled data can increase the model's performance and generalization.

Semi-supervised learning in speech analysis can be used for sentiment analysis, spam detection, document classification, training a model on a larger unlabelled text corpus and a smaller labelled text dataset, speech recognition using a large amount of unlabelled speech data and a small amount of transcribed speech data, etc. Semi-supervised learning uses a variety of algorithms, including generic semi-supervised algorithms, algorithms created specifically for semi-supervised learning, and conventional techniques that can be utilized by adding unlabelled data to the dataset.

3.2.4 Reinforcement Learning

Reinforcement learning uses training agents to make a series of decisions in an environment in order to maximize the reward signal. The AI agent (software component) investigates the environment by following, hitting, and acting; this allows it to gain knowledge from the experience and improve performance. These techniques use algorithms that have been taught through numerous trial-and-error experiments, learn from interactions in the environment, and get feedback in the form of incentives or penalties. Since the agent is rewarded for good behaviour and penalized for bad behaviour, reinforcement learning aims to maximize reward. It operates on agents' experiences, which are comparable to people's thinking.

The reinforcement learning approach is used in robotics to teach robots to fly, walk, or manipulate objects; autonomous driving to create self-driving cars; video games and games like chess; natural language processing (NLP) for translation, machine translation, question answering, text summarization, and dialogue generation; and more. Reinforcement learning employs three algorithms: Q-learning, Proximality Policy Optimization (PPO), and Deep Q Networks (DQN). This approach works well for sequential decision-making challenges, situations in which decisions have delayed effects, and environments with shifting dynamics and uncertainty.

3.3 Exploratory Data Analysis (EDA)

EDA is a statistical method for comprehending the properties of a given dataset that makes use of data visualizations. John Tukey popularized the EDA method in 1970, which aids in both hypothesis building and data exploration. It is an analysis method used to learn more about the dataset, including a synopsis of its key features, and it offers data visualization following analysis. EDA assists in collecting unexpected data, choosing the appropriate statistical methods and tools, eliminating extraneous values and irregularities, and lowering the likelihood of errors in subsequent analysis. In EDA, quantile-quantile (Q-Q) plots, boxplots, histograms, and other data visualization techniques are used to comprehend the distribution of data.

3.4 Deep Learning

Deep Learning is a potent subset of machine learning that models intricate patterns in data by simulating the structure and learning process of the human brain using artificial neural networks. Multiple neuronal layers, activation functions like ReLU and Sigmoid, and optimization strategies like gradient descent and backpropagation are also involved. Among the deep learning models are Transformers for natural language processing applications, Feedforward Neural Networks (FNNs) for general tasks, Convolutional Neural Networks (CNNs) for image processing, and Recurrent Neural Networks (RNNs) for sequential data. Deep learning is extensively utilized in the healthcare industry for drug discovery, disease prediction, medical imaging, and patient monitoring through Electronic Health Records (EHR). Deep learning makes it possible to create AI-driven models that increase diagnostic precision, streamline medical procedures, and improve patient outcomes using frameworks like TensorFlow and Py-Torch.

3.4.1 Fully Connected Neural Network (FCNN)

Every neuron in one layer of an artificial neural network (ANN) is connected to every other neuron in the layer below it, forming a fully connected neural network (FCNN). An input layer, several hidden layers, and an output layer make up this system, which adds non-linearity using activation functions as ReLU, Sigmoid, or Softmax. When working with large datasets, FCNNs can become computationally costly and prone to overfitting, despite their widespread use in classification, regression, and feature extraction. Even so, they are essential to deep learning applications and act as the foundation for more intricate architectures.

3.4.2 Wide and Deep Neural Network (WDNN)

A wide (shallow) network and a deep neural network are combined in a hybrid deep learning model called a Wide and Deep Neural Network (WDNN) to increase learning efficiency. In contrast to the deep network (deep neural layers), which learns more abstract and complicated representations, the wide network (linear model) records the memorization of past patterns and feature interactions. Because they strike a balance between short-term relevance (memorization) and long-term generalization, WDNNs are very useful in recommendation systems, predictive analytics, and personalized search. Businesses like Google use this technique extensively for their ranking and app recommendation algorithms.

3.5 Principal Component Analysis (PCA)

An unsupervised machine learning method called Principal Component Analysis (PCA) is utilized to reduce dimensionality while keeping as much crucial information as feasible. By locating the most important patterns, or primary components, it converts high-dimensional data into a lower-dimensional space.

Standardization is the first step in the process, in which the data is scaled to have a variance of 1 and a mean of 0. The associations between various features are then determined by calculating a covariance matrix. The new axes (principal components) that best explain the data's variance are then identified by computing the eigenvectors and eigenvalues. In order to minimize information loss and reduce dimensionality, the data is finally projected onto these new primary components. Feature selection, data reduction, visualization, and enhancing machine learning model performance by eliminating superfluous features are all common applications for PCA.

3.6 K-Means Clustering

K-Means Clustering is an unsupervised machine learning approach that minimizes the distance between data points and their designated cluster centroids in order to arrange data points into K different clusters based on feature similarity.

In order to determine how many clusters to generate, the user must first select the number of clusters (K). The centroids are then initialized by choosing K sites at random to serve as the first cluster centre. Next, using Euclidean distance, each data point is matched with the closest centroid. The mean of each cluster's points is then calculated to update the centroids. In order to ensure convergence, these assignment and update procedures are repeated until the centroids cease changing or a stopping requirement is satisfied.

3.7 Classification Models

The machine learning algorithm includes a number of classification models that assist in grouping the dataset into distinct classes according to their traits when taught using training data and tested using test data that is not visible. Support Vector Machine (SVM), Random Forest, Logistic Regression, and gradient boosting are some of the categorization models that were employed in this investigation.

3.7.1 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a machine learning system that uses supervised learning for both classification and regression problems. Creating the decision border that divides various classes in n-dimensional space is the aim of SVM. Because the support vectors or extreme points aid in the creation of the hyperplane, the technique is known as a Support Vector Machine.

SVM determines the optimal border or hyperplane in an N-dimensional space that divides data points into distinct classes in linear classification tasks. The maximum margin hyperplane, also known as the ideal hyperplane, is the greatest separation between the boundary and the nearest data point in each class. The support vectors are the data points that are nearest to each other. The kernel trick is used to make non-linearly separable data linearly separable by mapping the data into higher-dimensional space. The hyperplane is defined by an orthogonal collection of vectors, where the set of vectors in that space and the dot product of data points are constants.

3.7.2 Logistic regression

Logistic regression is a machine learning algorithm that comes under the supervised learning technique for solving classification problems. It predicts categorical dependent variables using the specified set of independent variables. Results will be in the form of true or false, 0 or 1, or categorical values such as with disease or without disease. In order to provide a single numerical value, logistic regression first computes a weighted sum of the input features plus a bias component. The sigmoid activation function is then applied to this number, converting it into a probability between 0 and 1. The instance is assigned to the positive class (1) if the probability is greater than a predetermined threshold, such 0.5; if not, it is assigned to the negative class (0). Because of this procedure, logistic regression is a popular classification algorithm that can successfully differentiate between two categories.

3.5.3 Random Forest

The most well-liked machine learning algorithm for supervised learning techniques, Random Forest, can be applied to both regression and classification issues. It is a classifier that uses many decision trees on various dataset subsets. Each decision tree makes predictions, and the final output is determined by calculating the majority of the predictions. An increased number of trees in the forest will increase accuracy and guard against overfitting. Assume that each tree's predictions have less correlations for a better Random Forest classifier, and that the dataset's feature variable contains some real values. Consequently, a classifier can forecast precise outcomes rather than relying on conjecture.

The random forest process consists of building N decision trees from the provided dataset using K randomly chosen data points from the training dataset, then feeding the model with a fresh, unknown dataset, also known as the test dataset, from which each decision tree generates predictions. The category with the majority of votes will be given to test data points based on these predictions.

3.7.4 Gradient boosting

Gradient boosting is a supervised machine learning algorithm an ensemble learning technique that can be used for both classification and regression tasks. It creates a sequence of weak learners (usually decision trees) one after the other, correcting the mistakes of the preceding model. By iteratively learning from errors, gradient boosting lessens bias in contrast to bagging techniques like Random Forest.

First, a weak model is used to make an initial prediction. Next, the residual errors the difference between the actual and projected values are calculated. After that, a new model is trained to forecast these residuals, and its predictions are weighted and added gradually to those of the prior model. Until the error is reduced, this iterative process keeps going. Sequential learning, bias and variance reduction, and the capacity to manage missing data efficiently are some of Gradient Boosting's primary characteristics. High prediction accuracy, robust performance on structured/tabular data, and the capacity to effectively manage noisy data and missing values.

3.8 Correlation Matrix

A correlation matrix is a table that displays the correlation coefficients between multiple variables in a dataset. It helps to understand the relationships between variables by measuring how strongly they are related to each other. The correlation values range from -1 to 1:

+1 (Perfect Positive Correlation) → As one variable increases, the other also increases.

0 (No Correlation) → No relationship between the variables.

-1 (Perfect Negative Correlation) → As one variable increases, the other decreases.

A correlation matrix is useful because it helps identify strong and weak relationships between variables, making it easier to understand dependencies in a dataset. It is particularly valuable in feature selection for machine learning models, as highly correlated variables may provide redundant information. Additionally, it helps detect multicollinearity, a condition where two or more independent variables are highly correlated, which can negatively impact model performance by making coefficient estimates unstable

3.9 Model Evaluation

Model evaluation is the process of assessing the performance of a machine learning model to determine how well it generalizes to new, unseen data. It involves using various metrics and techniques to measure the accuracy, efficiency, and reliability of the model. Common evaluation metrics depend on the type of task: for classification models, metrics like accuracy, precision, recall, F1-score, and ROC-AUC are used, while for regression models, metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) are commonly applied. Cross-validation techniques like K-Fold Cross-Validation help ensure that the model performs consistently across different subsets of the data, reducing the risk of overfitting or underfitting. Proper model evaluation is essential to selecting the best model and optimizing its parameters for better real-world performance.

3.9.1 Confusion Matrix

The confusion matrix is used for determining the performance of the classification models in the given test data. Performance can be determined if there are true values for the test data. Also called an error matrix since the errors in model performance are displayed in the form of a matrix. The matrix has two dimensions one is predicted values and the other is actual values with the total number of predictions. Actual values are the true values for the given observations and predicted values are the values predicted by the model.

The confusion matrix for this study can be interpreted as:

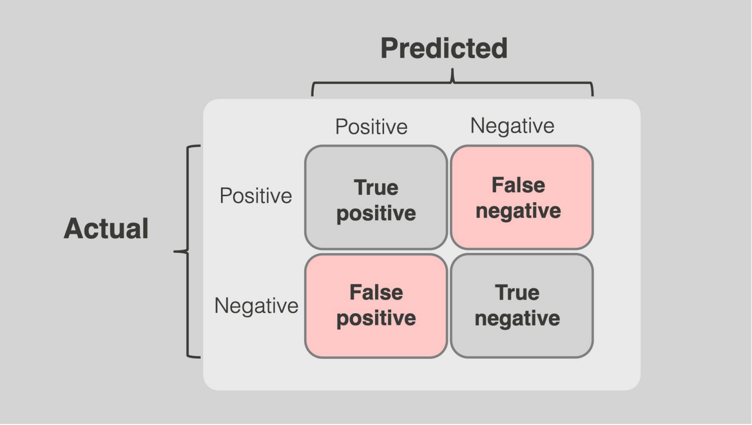


Figure:3.1 Interpretation of confusion matrix

True Positives (TP) – Correctly predicted positive cases.

True Negatives (TN) – Correctly predicted negative cases.

False Positives (FP) – Incorrectly predicted positives (Type I error).

False Negatives (FN) – Incorrectly predicted negatives (Type II error)

Various calculations are done using this confusion matrix for evaluating the performance of the model such as:

Accuracy: This determines the accuracy of the classification problem and defines how the model predicts the correct output is the number of correct predictions to all the number of predictions made by the model.

$$\text{Accuracy} = \frac{TN+TP}{TN + FP+FN+TP}$$

Precision: The number of actual true predictions to all the positive class predictions made by the model.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall: It is defined as how the model predicted correctly out of total predictions.

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-Score: It is difficult to compare the models if has high recall and low precision or vice versa. In this situation evaluation of recall and precision is done at the same time using F1-score. If recall is equal to precision, then the F1-score is maximum.

$$F1 - score = \frac{2 * Recall * precision}{Recall + Precision}$$

3.9.2 ROC-AUC Curve

The receiver operating characteristic (ROC) curve is a graphical plot that plots the varying threshold values and depicts the performances of a binary classifier or multi-class classifier. This plots the False Positive Rate (FPR) against the True Positive Rate (TPR) at each threshold. Plots the statistical power as a function of Type-II Error in the decision rule thus, this curve is sensitivity or recall as a function of False Positive rate. If the probability distribution of False Positive and True Positive is known then the Cumulative Distribution Function (CDF) is shown for the ROC curve. From the ROC curve, we can identify the optimal model independently from the class distribution or the class context.

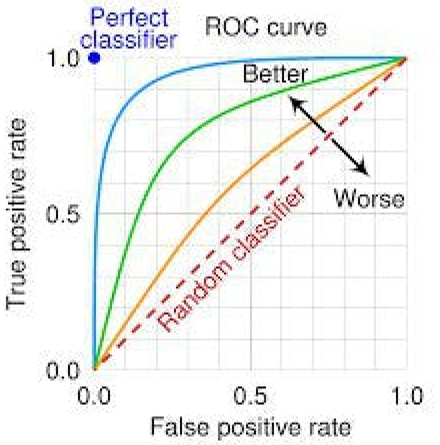


Figure 3.2 ROC -AUC Curve

The plot of a ROC curve is shown in Figure 3.2, to portray this only the True Positive Rate (TPR) and the False Positive Rate (FPR) are needed. Plotting FPR along the x-axis and TPR along the y-axis demonstrates the relative trade-off between true positives and false positives. The ROC curve is sometimes called the sensitivity vs specificity plot, where TPR represents sensitivity and FPR represents specificity. Each point in the ROC space is the prediction result of the confusion matrix.

When there is a point in the ROC space with coordinate (0,1) this is the best prediction method possible called perfect classification. The line through the diagonal of the ROC space connecting the bottom left corner and the top right corner is the line of no discrimination. This diagonal divides the space into better and worse classification spaces. The prediction results above the diagonal show better classification and those below the diagonal show worse classification.

3.10 Feature Importance

A feature importance plot is a graphical plot that exemplifies the significance of different variables in a dataset in predicting outputs using a machine learning model. This plot commonly used in data analytics, statistics, and machine learning helps in understanding which features in the given dataset have relative importance in the model's prediction. Feature importance plot helps in gaining insights about the relationship between features and the target variables and also for explaining the results of a model.

3.11 Methodology of the study

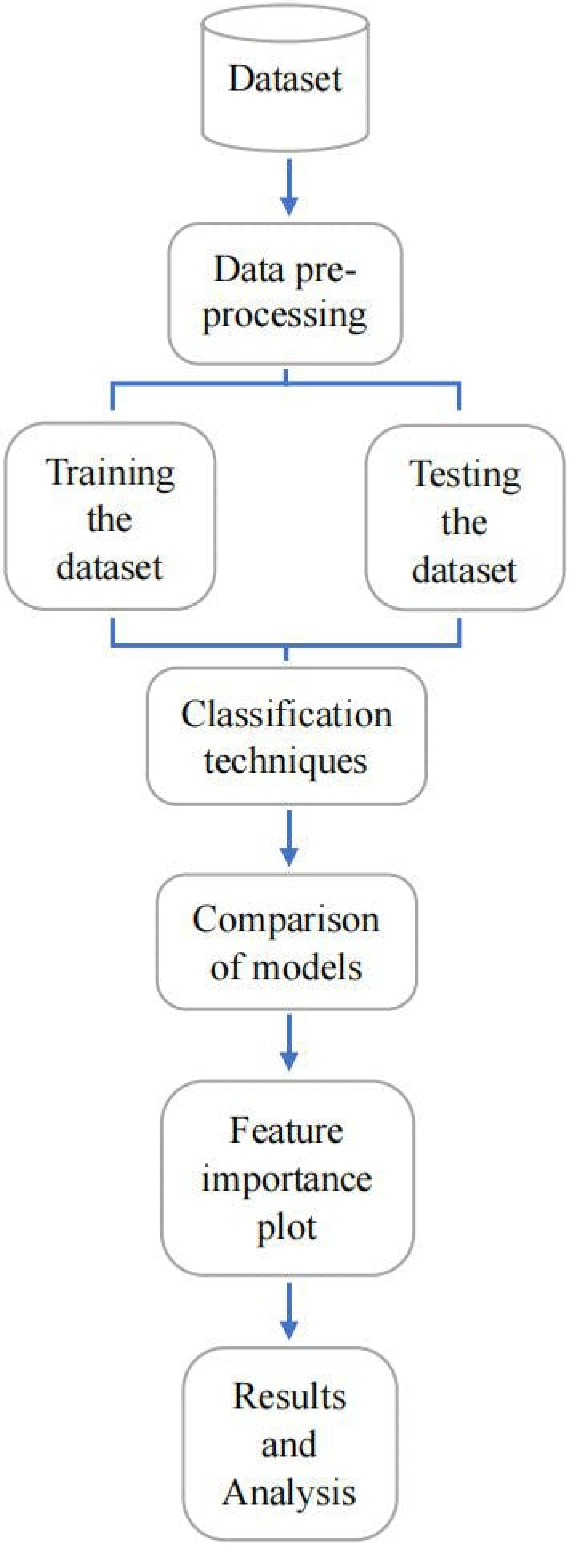


Figure 3.3 Methodology of the study

Chapter 4

Data Description and Exploratory Data Analysis

4.1 Attributes of the dataset

The Liver Disease Patient dataset, sourced from open-access website Kaggle, consists of 30,691 patient records collected from Electronic Health Records (EHR), having 10 attributes making it a valuable resource for medical research and predictive modelling. The dataset includes demographic details such as age and gender, along with clinical biomarkers like total and direct bilirubin, alkaline phosphate, aspartate aminotransferase (SGOT), alanine aminotransferase (SGPT), albumin, and the albumin-to-globulin ratio

Dataset Structure:

Demographic Features:

i. Age of the patient
ii. Gender of the patients

Liver Function Test Parameters:

1. Total Bilirubin – A yellow pigment produced during the breakdown of red blood cells. Elevated levels may indicate liver disease or bile duct obstruction.

2. Direct Bilirubin – The conjugated form of bilirubin processed by the liver. High levels suggest liver dysfunction or bile flow obstruction.

3. Alkaline Phosphatase (Alkphos) – An enzyme found in the liver, bones, and other tissues. Increased levels may indicate liver or bone disorders.

4. Alamine Aminotransferase (SGPT/ALT) – An enzyme involved in liver metabolism. High ALT levels can be a marker of liver damage or inflammation.

5. Aspartate Aminotransferase (SGOT/AST) – An enzyme found in the liver, heart, and muscles. Elevated levels may indicate liver disease or heart conditions.

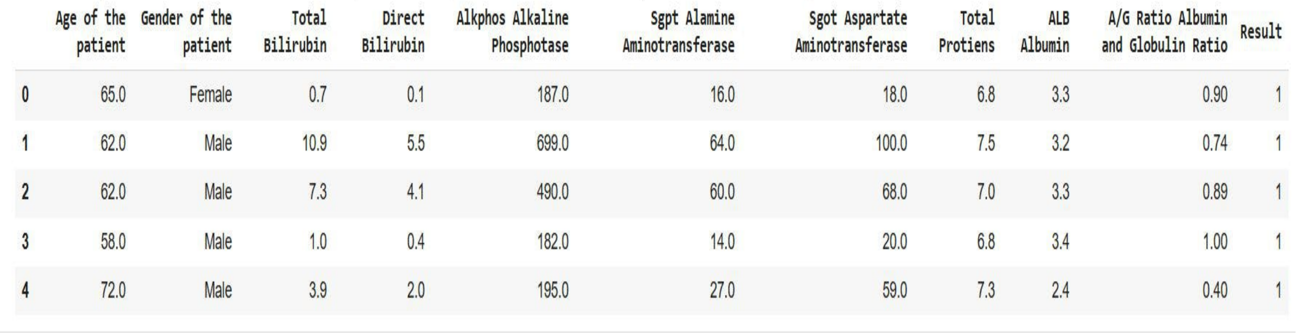
6. Total Proteins – The sum of all proteins in the blood, including albumin and globulins. Low levels may indicate liver or kidney disease.

7. Albumin (ALB) – A major protein produced by the liver, responsible for maintaining blood volume and transporting substances. Low levels may indicate liver disease.

8. Albumin and Globulin Ratio (A/G Ratio) – The ratio between albumin and globulin proteins. A low A/G ratio can indicate chronic liver disease or other medical conditions.

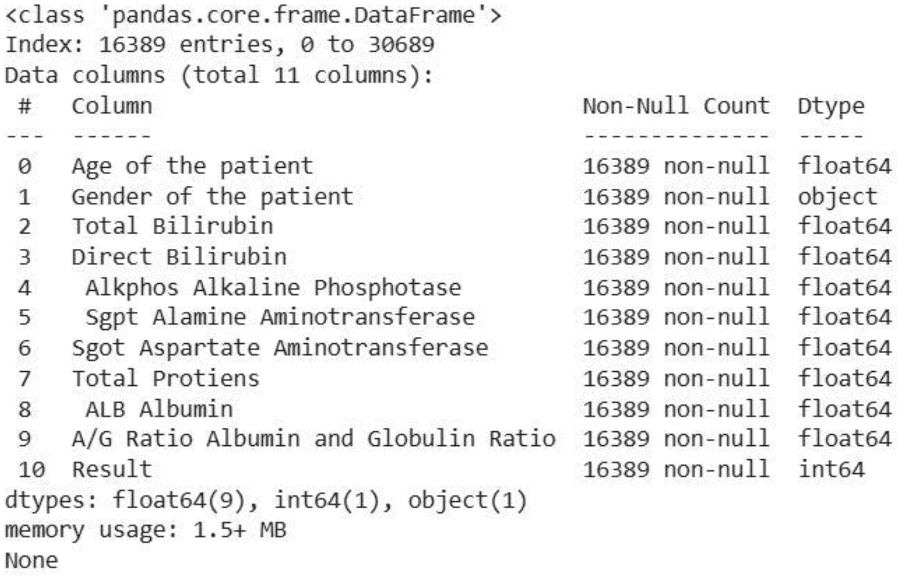
The dataset divides into patients with liver disease (21,917cases) and patients without liver disease (8,774 cases) uses 1 for "Disease Present" and 2 for "No Disease". The aim of the study. The main aim of this study is to diagnose whether the sample is with disease or without disease. Using models for the prediction of the disease or no disease from the dataset.

4.2 Sample Dataset



	Age of the patient	Gender of the patient	Total Bilirubin	Direct Bilirubin	Alkphos Alkaline Phosphotase	Sgpt Alamine Aminotransferase	Sgot Aspartate Aminotransferase	Total Protiens	ALB Albumin	A/G Ratio Albumin and Globulin Ratio	Result
0	65.0	Female	0.7	0.1	187.0	16.0	18.0	6.8	3.3	0.90	1
1	62.0	Male	10.9	5.5	699.0	64.0	100.0	7.5	3.2	0.74	1
2	62.0	Male	7.3	4.1	490.0	60.0	68.0	7.0	3.3	0.89	1
3	58.0	Male	1.0	0.4	182.0	14.0	20.0	6.8	3.4	1.00	1
4	72.0	Male	3.9	2.0	195.0	27.0	59.0	7.3	2.4	0.40	1

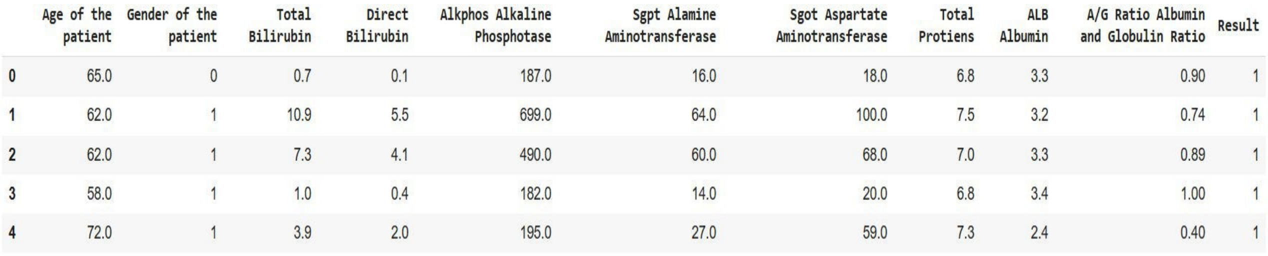
Figure 4.1 Head of the dataset



```
<class 'pandas.core.frame.DataFrame'>
Index: 16389 entries, 0 to 30689
Data columns (total 11 columns):
 #   Column                               Non-Null Count  Dtype  
---  ------                               --------------  -----  
 0   Age of the patient                   16389 non-null  float64
 1   Gender of the patient                16389 non-null  object 
 2   Total Bilirubin                      16389 non-null  float64
 3   Direct Bilirubin                     16389 non-null  float64
 4    Alkphos Alkaline Phosphotase        16389 non-null  float64
 5    Sgpt Alamine Aminotransferase       16389 non-null  float64
 6   Sgot Aspartate Aminotransferase      16389 non-null  float64
 7   Total Protiens                       16389 non-null  float64
 8    ALB Albumin                         16389 non-null  float64
 9   A/G Ratio Albumin and Globulin Ratio 16389 non-null  float64
 10  Result                               16389 non-null  int64  
dtypes: float64(9), int64(1), object(1)
memory usage: 1.5+ MB
None
```

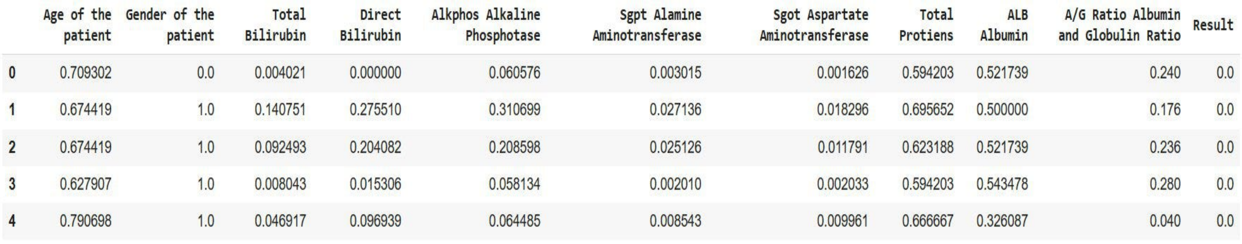
Figure 4.2 Handling the dataset

Figure 4.1 is the head of the original dataset further after handling the missing values and the replicated data are removed for better performance and accuracy. After removing the total no of patient record is 16,389 from 30,689 records. Then the dataset is set convert the categorical to numerical value Figure4.3 and forwarded with normalisation Figure 4.4.



	Age of the patient	Gender of the patient	Total Bilirubin	Direct Bilirubin	Alkphos Alkaline Phosphotase	Sgpt Alamine Aminotransferase	Sgot Aspartate Aminotransferase	Total Protiens	ALB Albumin	A/G Ratio Albumin and Globulin Ratio	Result
0	65.0	0	0.7	0.1	187.0	16.0	18.0	6.8	3.3	0.90	1
1	62.0	1	10.9	5.5	699.0	64.0	100.0	7.5	3.2	0.74	1
2	62.0	1	7.3	4.1	490.0	60.0	68.0	7.0	3.3	0.89	1
3	58.0	1	1.0	0.4	182.0	14.0	20.0	6.8	3.4	1.00	1
4	72.0	1	3.9	2.0	195.0	27.0	59.0	7.3	2.4	0.40	1

Figure 4.3 converted to numerical value



	Age of the patient	Gender of the patient	Total Bilirubin	Direct Bilirubin	Alkphos Alkaline Phosphotase	Sgpt Alamine Aminotransferase	Sgot Aspartate Aminotransferase	Total Protiens	ALB Albumin	A/G Ratio Albumin and Globulin Ratio	Result
0	0.709302	0.0	0.004021	0.000000	0.060576	0.003015	0.001626	0.594203	0.521739	0.240	0.0
1	0.674419	1.0	0.140751	0.275510	0.310699	0.027136	0.018296	0.695652	0.500000	0.176	0.0
2	0.674419	1.0	0.092493	0.204082	0.208598	0.025126	0.011791	0.623188	0.521739	0.236	0.0
3	0.627907	1.0	0.008043	0.015306	0.058134	0.002010	0.002033	0.594203	0.543478	0.280	0.0
4	0.790698	1.0	0.046917	0.096939	0.064485	0.008543	0.009961	0.666667	0.326087	0.040	0.0

Figure 4.4 Normalised dataset

4.3 Exploratory Data Analysis

Conducting exploratory data analysis for ideas regarding the dataset from some of the basic results obtained.

4.3.1Bar plot

A **bar plot** is a graphical representation of categorical data using rectangular bars, where the height or length of each bar corresponds to the value of the category. It is commonly used to compare different groups, visualize distributions, and identify trends in data.

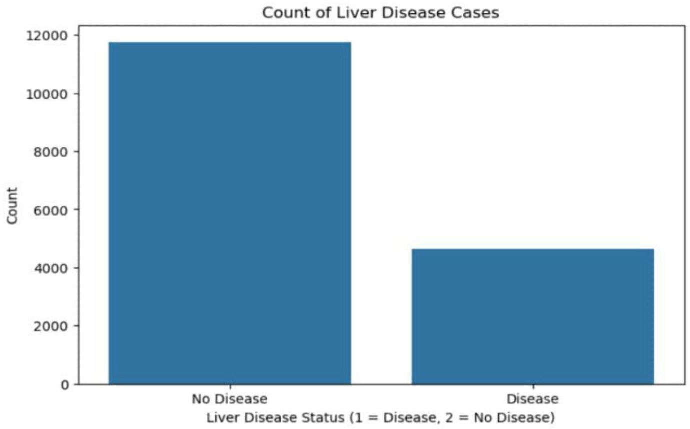


Figure 4.5 Count of liver diseases cases

From the Figure 4.5

The **"No Disease"** group has a significantly higher count (around12,00 cases). The **"Disease"** group has fewer cases (around 5,000).

4.3.2 Kernel Density Estimation (KDE) plot

A Kernel Density Estimation (KDE) plot is a smooth representation of a histogram, used to estimate the probability density function (PDF) of a continuous variable. It helps visualize the distribution of data without being affected by bin sizes (like histograms).

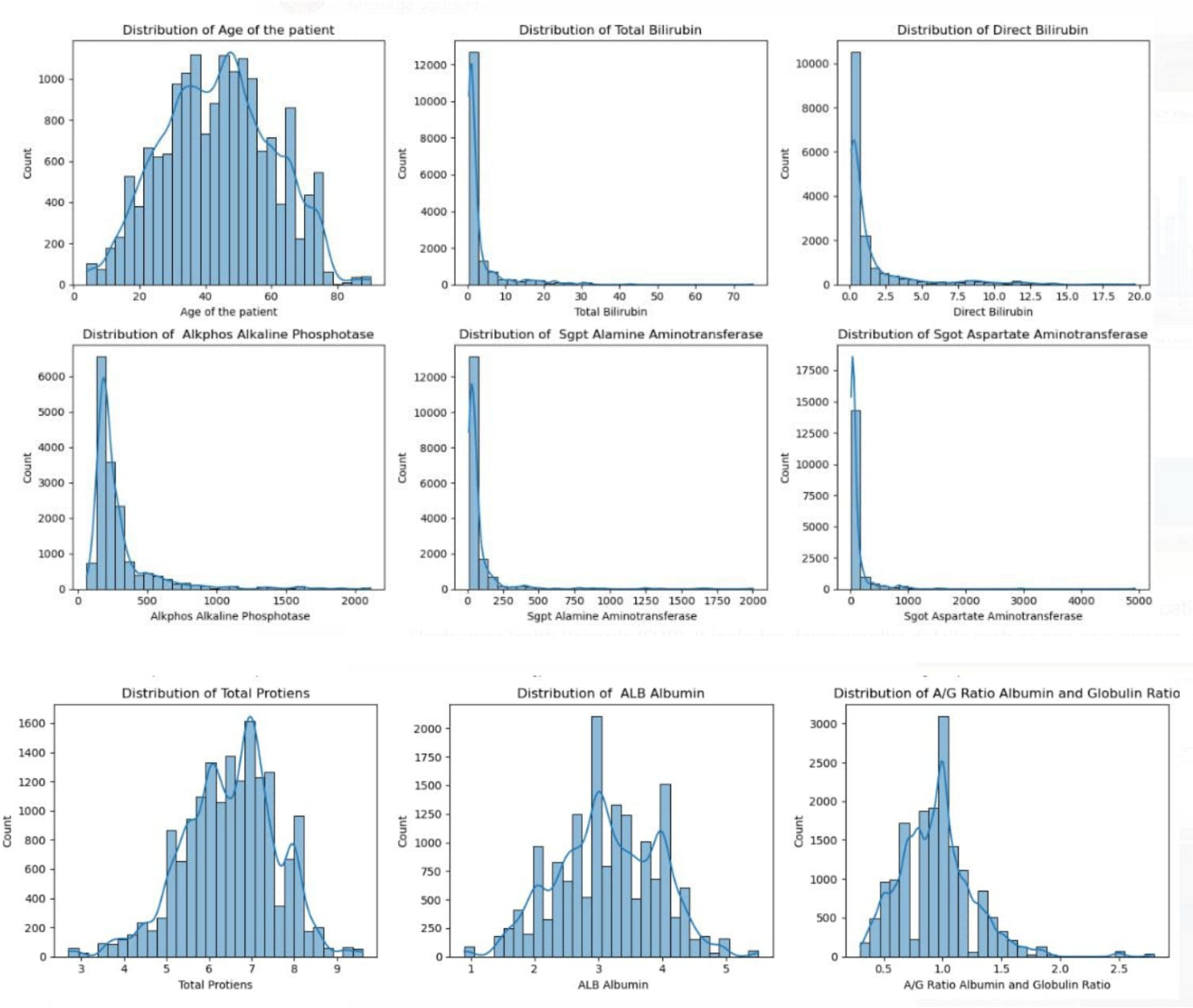


Figure 4.6 KDE Plot

Histograms (bars) show the distribution of numerical data by grouping values into bins and counting occurrences.
KDE curves (smooth lines) estimate the probability density function of the data, providing a smoothed view of the distribution.

4.3.3 Box Plot

A box plot (or box-and-whisker plot) is a statistical visualization that represents the distribution, variability, and potential outliers in a dataset. It consists of a median (Q2), shown as a line inside the box, and the interquartile range (IQR), which spans from Q1 (25th percentile) to Q3 (75th percentile), covering the middle 50% of the data. Whiskers extend beyond the box, typically up to 1.5 × IQR from Q1 and Q3, indicating the range of the data. Outliers appear as individual points beyond the whiskers, representing extreme values in the dataset.

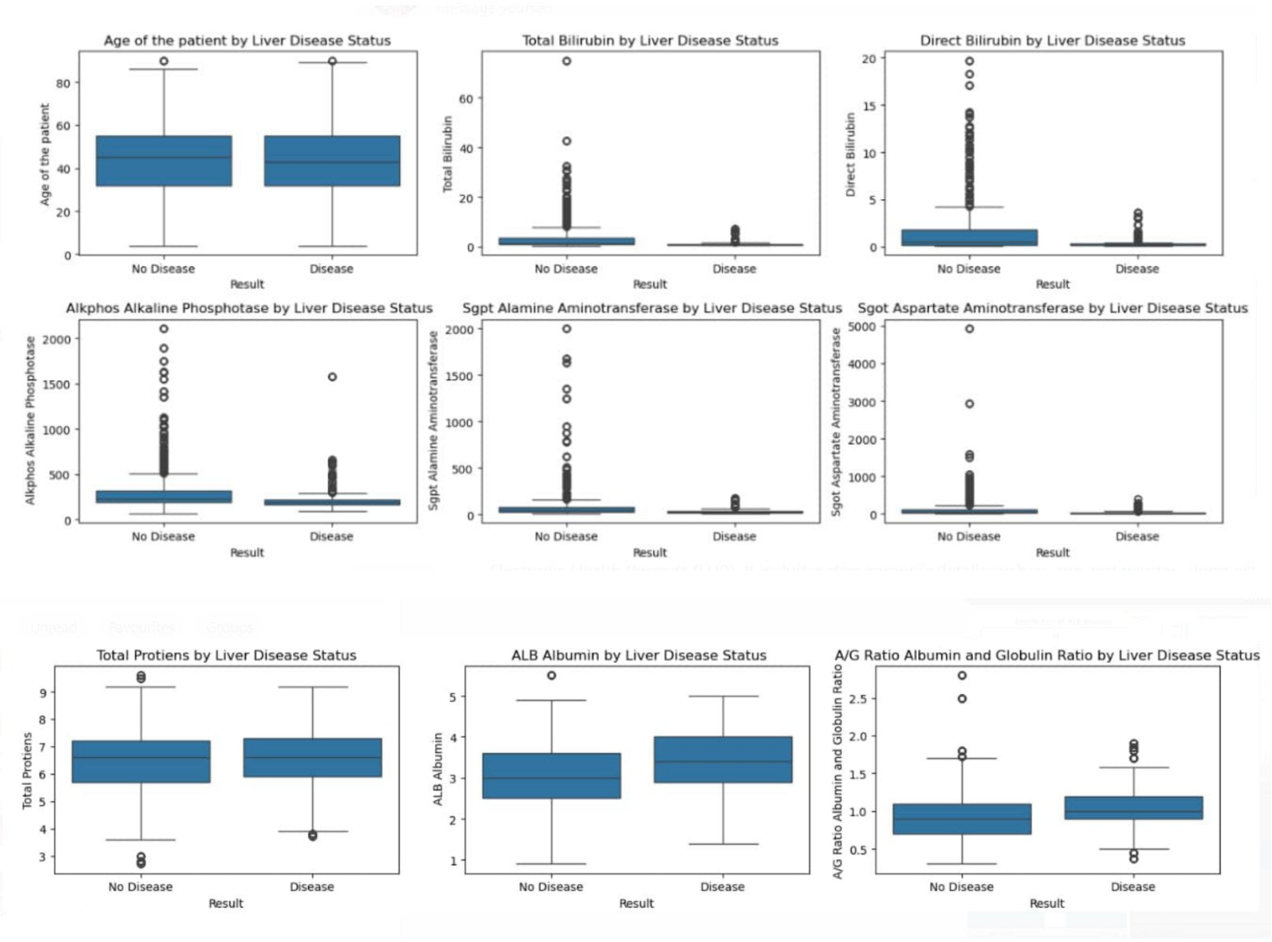


Figure 4.7 Box Plot

4.3.4 Heat Map

A heatmap is a data visualization technique that represents values using colours, making it easy to identify patterns, relationships, and intensity variations in data. It uses a colour gradient to represent numerical values, with darker colours indicating higher values and lighter colours indicating lower values. Heatmaps are often used for correlation matrices to display relationships between variables and can also highlight missing values, outliers, or clusters within a dataset.

Findings:

1. Strongly Correlated Features: Some features, like Total Bilirubin, Direct Bilirubin, Sgpt, and Sgot, are strongly correlated, indicating their joint significance in liver health assessment.
2. Weak Predictors for Result: Features like Age and Gender are weakly correlated with the target variable, meaning they may not play a significant role in predicting liver disease.
3. Key Features: Features like Direct Bilirubin, Total Bilirubin, ALB Albumin, and A/G Ratio show a stronger relationship with the target variable.

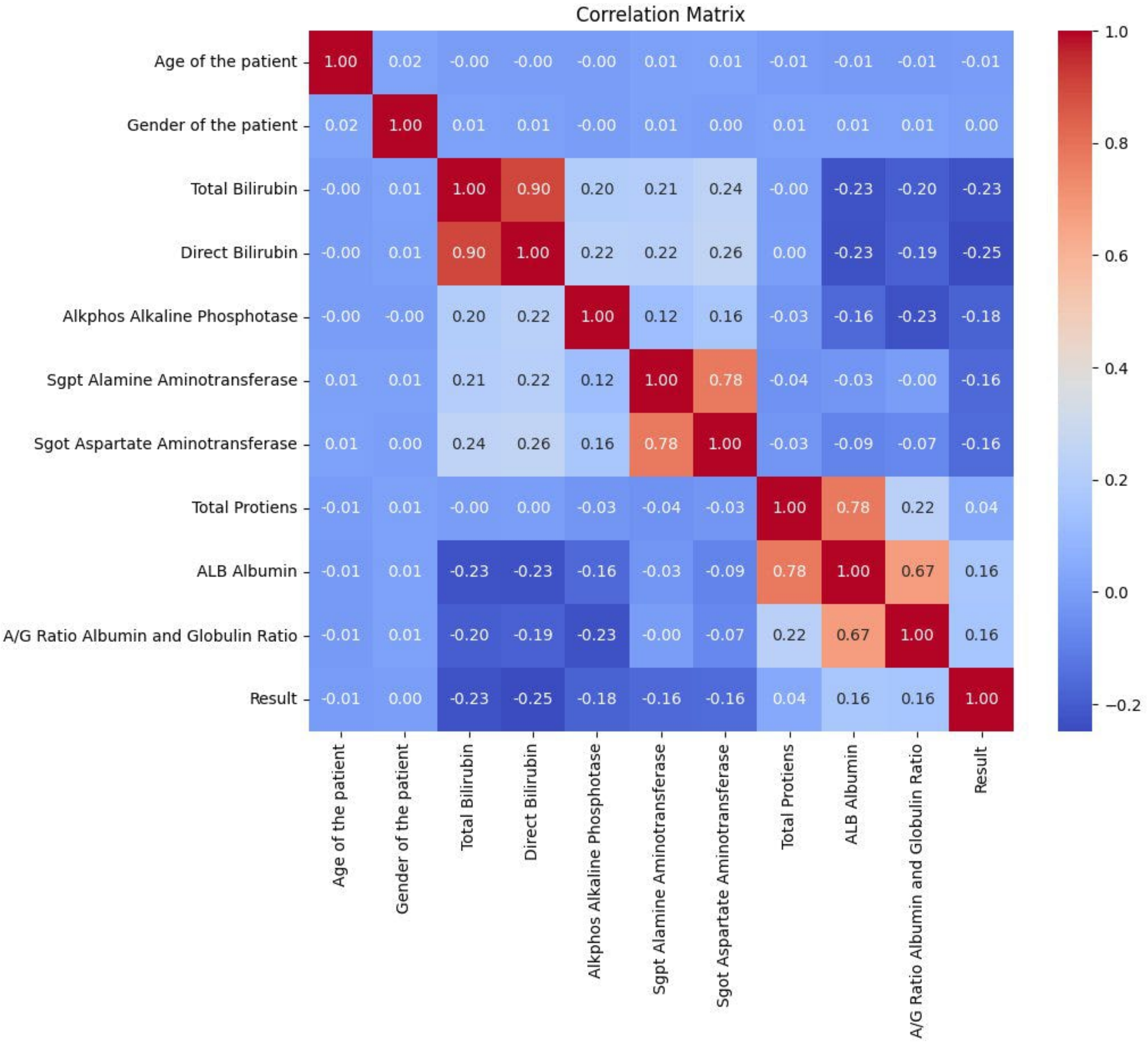


Figure 4.8 Heat Map

Chapter 5

Analysis and Results

5.1 Dimensionality reduction using Principal Component Analysis (PCA)

5.1.1 Dimensionality reduction

Using principal component analysis (PCA) with a 2 - component projection Figure 5.1 retaining approximately of the total variance, to reduce dimensionality and enhance feature interpretability."

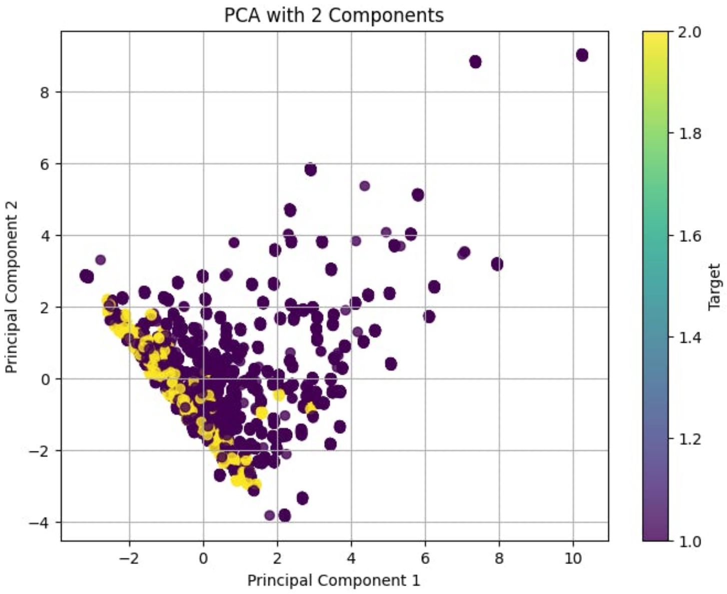


Figure 5.1 PCA n=2

Explained Variance Ratio:

PC1: 0.27, PC2: 0.19

Principal Component Loadings:

	PC 1	PC2
Age of the patient	0.008213	-0.007792
Gender of the patient	0.000893	0.024030
Total Bilirubin	0.441942	0.224196
Direct Bilirubin	0.445111	0.232110
AlkphosAlkaline Phosphotase	0.251765	0.034425
Sgpt Alamine Aminotransferase	0.289050	0.393376

Sgot Aspartate Aminotransferase	0.324649	0.375707
Total Protien	-0.250814	0.479044
ALB Albumin	-0.414731	0.497207
A/G Ratio	-0.345331	0.348481

Table 5.1 PC loadings

Top Contributing Features for PC1

Direct Bilirubin	0.445111
Total Bilirubin	0.441942
ALB Albumin	0.414731

Table 5.2 Features for PC 1

Top Contributing Features for PC2:

ALB Albumin	0.497207
Total Protiens	0.479044
Sgpt Alamine Aminotransferase	0.393376

Table 5.3 Features for PC 2

5.1.2 "PCA-Enhanced K-Means Clustering with Logistic Regression Evaluation"

The dataset has been reduced to two dimensions using Principal Component Analysis (PCA), represented by the x-axis (Principal Component 1) and y-axis (Principal Component 2). The K-Means algorithm identified two distinct clusters, The clusters are vertically separated along the Principal Component 1 (PC1) axis. Cluster 1 (yellow) lies primarily on the left side of the graph, while Cluster 0 (purple) occupies the right side. The vertical alignment of data suggests that PC2 plays a lesser role in distinguishing clusters, with most variation explained by PC1 as shown in Figure 5.

Investigate the features contributing to PC1 to understand why the data separates this way PC1 is likely driven by the most dominant risk factors for liver disease by feature analysis

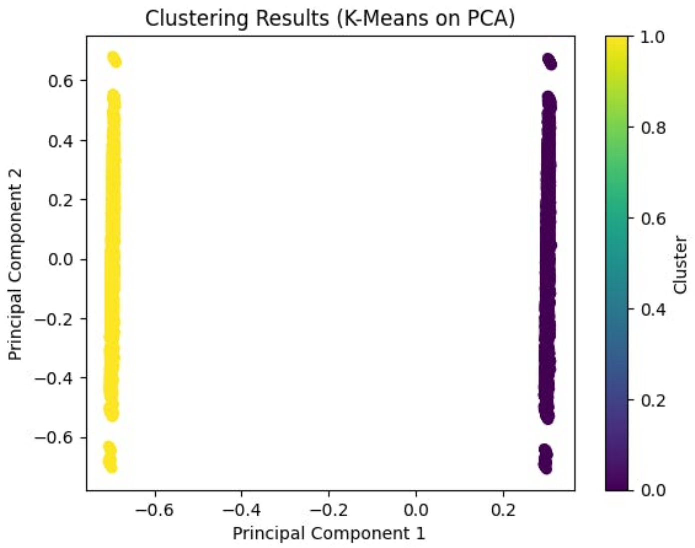


Figure 5.2 PCA-Enhanced K-Means Clustering with Logistic Regression Evaluation

Top Contributing Features for PC1:

ALB Albumin	0.468040
A/G Ratio Albumin and Globulin Ratio	0.367646
Total Protiens	0.326376
Gender of the patient	0.004396
Age of the patient	-0.011867
Sgpt Alamine Aminotransferase	-0.186026
Alkphos Alkaline Phosphotase	-0.209492
Sgot Aspartate Aminotransferase	-0.218671
Total Bilirubin	-0.343360
Direct Bilirubin	-0.345258
Cluster_Label	-0.419148

Table 5.4 Contributing Features for PC1:

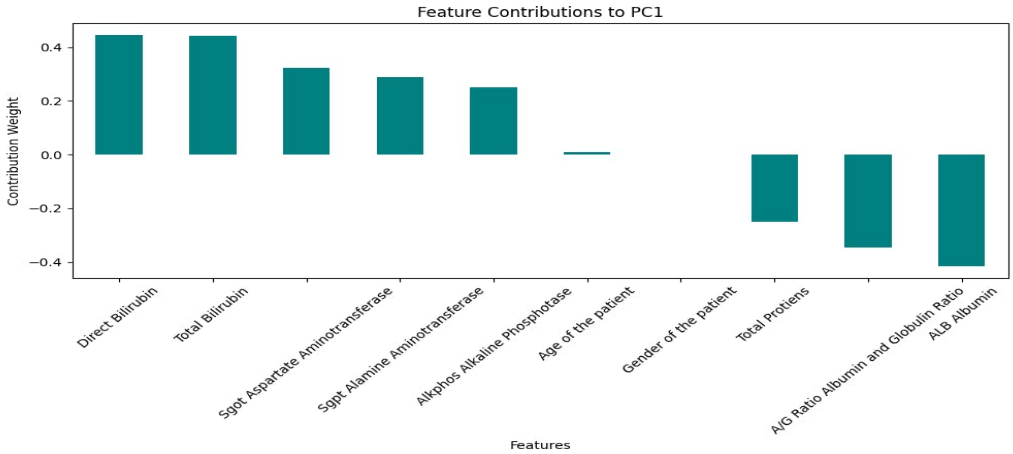


Figure 5.3 Feature contribution to PC 1

Further Evaluating using Logistic regression model gains accuracy of 0.71 percent on classification of 3515 and 0 into no disease cases and disease cases. 24 disease cases and 1378 no diseases cases were wrongly classified shown in the Figure 5.3. Further looking at other metrics such as precision, recall and F1 score are considered because of the classes are imbalanced Figure 5.3 and has an AUC value =0.63 depicted in Figure 5.4 of ROC curve

Accuracy: 0.71

ROC-AUC: 0.63

Classification table

	Precision	Recall	f1-score	Support
0	0.72	0.99	0.83	3539
1	0.76	0.68	0.71	1378
Accuracy			0.71	4917
Macro Average	0.36	0.50	0.41	4917
Weighted Average	0.52	0.71	0.60	4917

Table 5.5 Classification table of Logistic Regression

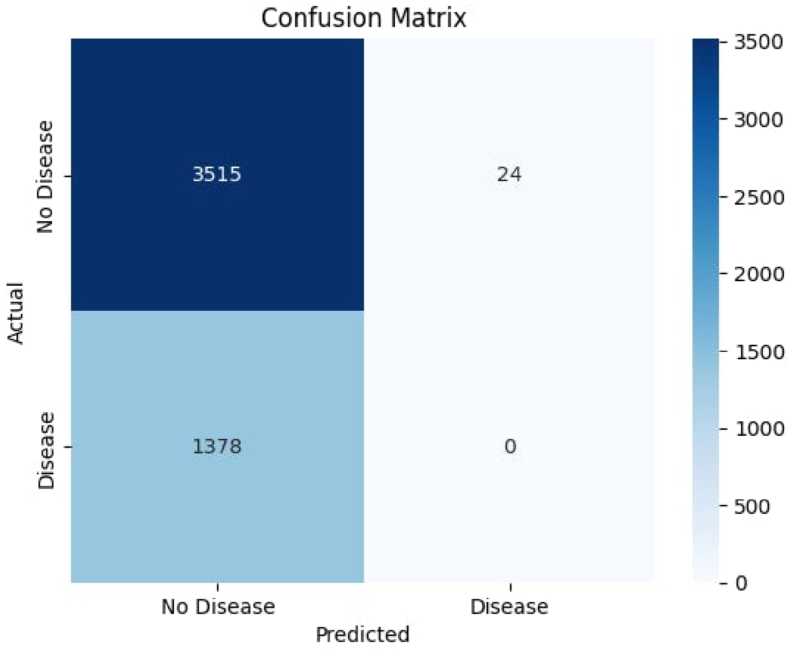


Figure 5.3 Confusion Matrix of Logistic regression

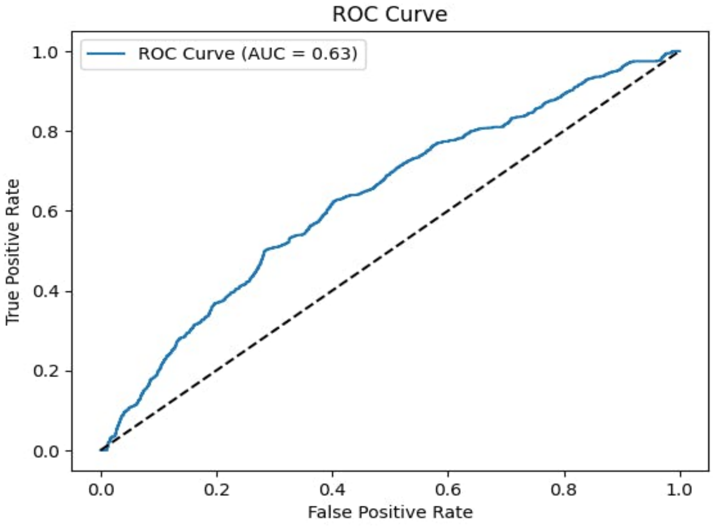


Figure5.4 ROC Curve of Logistic regression

5.1.3 PCA -Enhanced K-Means Clustering with Support Vector Machine (SVM)

These two components capture the most significant variance in the dataset after dimensionality reduction. From Figure5.5 the x-axis represents Principal Component 1 (PC1) and the y-axis represents Principal Component 2 (PC2). The red X marks represent the centroids of the clusters. Centroids are the mean positions of all points in their respective clusters. The data appears to be clustered into distinct groups, with clear divisions along PC1. This suggests that PC1 is the primary factor influencing the separation of clusters. One cluster (yellow) might represent cases with liver disease. Another cluster (purple) might represent cases without liver disease. The teal cluster might correspond to a specific subset of patients with distinct clinical or biological characteristics. It could represent cases with moderate risk factors or certain shared feature values (e.g., liver enzyme levels, age, etc.). Alternatively, it might include data points that are less related to the main patterns (e.g., cases with mixed or borderline outcomes).

Identify the most significant features driving Principal Component 1 (PC1) by Feature Analysis Figure 5.6

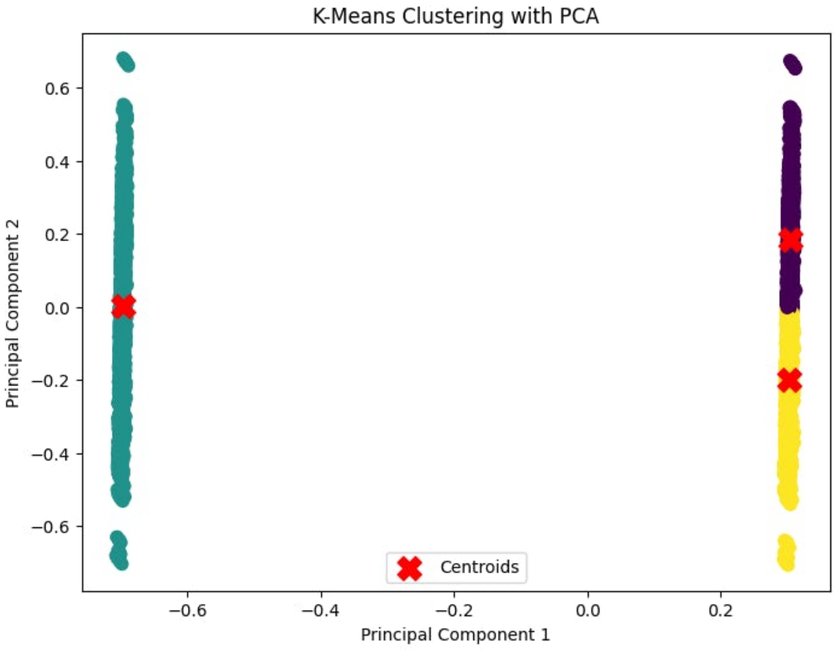


Figure 5.5 PCA -Enhanced K-Means Clustering with SVM

Top Contributing Features for PC1

Feature	Loading Absolute	Loading
Direct Bilirubin	0.445111	0.445111
Total Bilirubin	0.441942	0.441942
ALB Albumin -	0.414731	0.414731
A/G Ratio	-0.345331	0.345331
Sgot Aspartate Aminotransferase	0.324649	0.324649

Table 5.6 Contributing Features for PC1

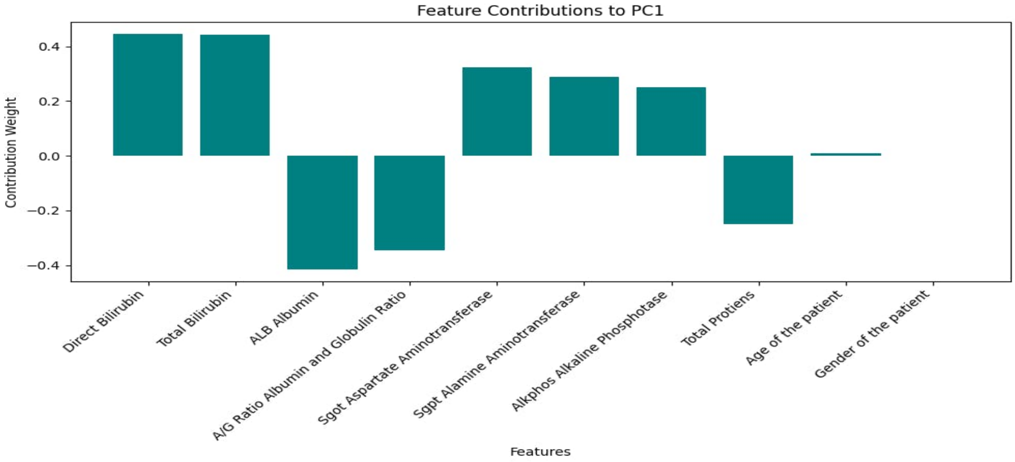


Figure 5.6 Feature contribution to PC 1

Analysing the Relationship Between Clusters and Target Variable

To understand how the clusters correspond to the presence or absence of liver disease

```
Cluster- Result Relationship:
RESULT    0.0   1.0
Cluster
    0       8182   3232
    1       3568   1407
```

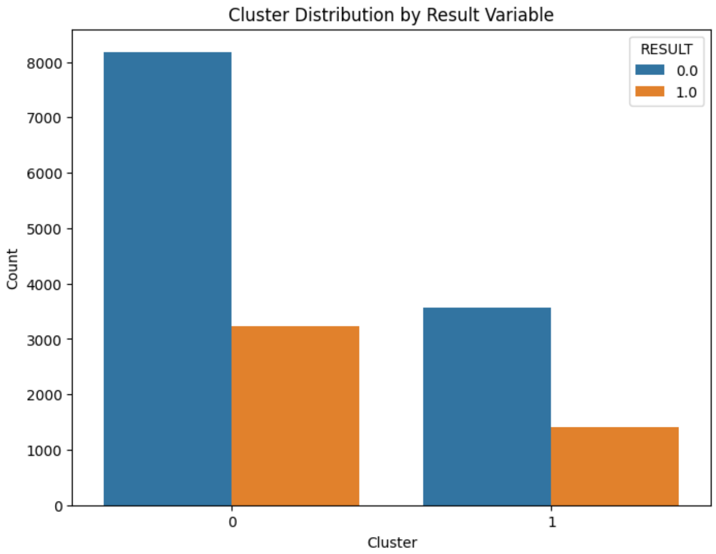


Figure 5.7 Cluster- Result Relationship

Further Evaluating using Support Vector Machine (SVM) model gains accuracy of 0.716 percent on classification of 3525 and 0 into no disease cases and disease cases. 0 disease cases and 1392 no diseases cases were wrongly classified shown in the Figure 5.3. Further looking at other metrics such as precision, recall and F1 score are considered because of the classes are imbalanced.

Accuracy: 0.71

ROC-AUC: 0.54

Classification table

	Precision	Recall	f1-score	Support
0	0.72	1.00	0.83	3539
1	0.00	0.00	0.00	1378
Accuracy			0.71	4917
Macro Average	0.36	0.50	0.42	4917
Weighted Average	0.52	0.72	0.60	4917

Table 5.7 Classification table of SVM

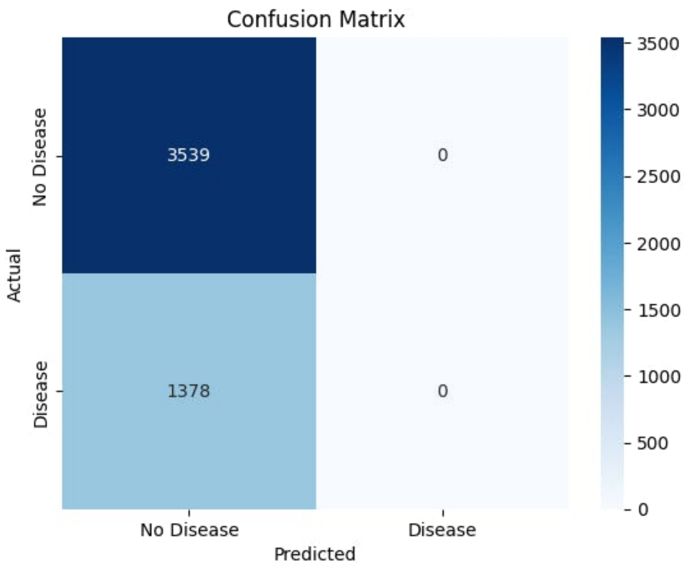


Figure 5.7 Confusion matrix of SVM

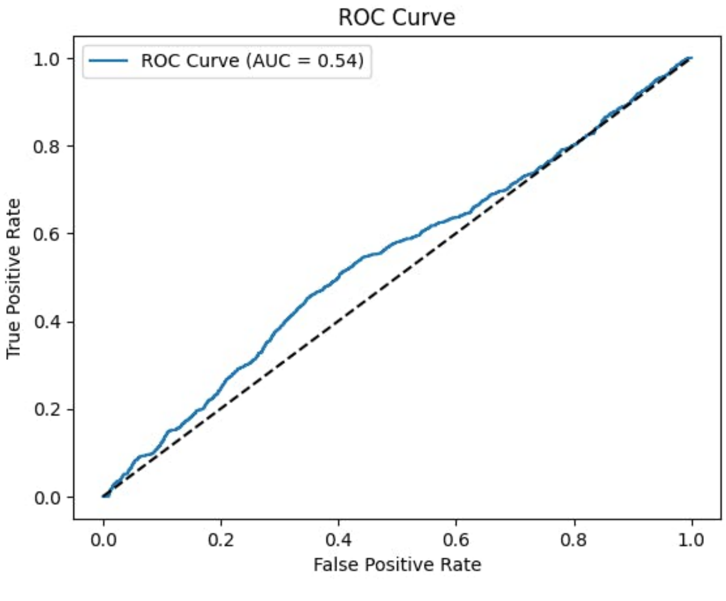


Figure 5.8 ROC Curve of SVM

5.2 ML Model Building

5.2.1 Gradient Boosting

The Gradient Boosting Model has accuracy of 89 percent in classifying 3462 of no disease cases and 895 of disease cases. Out of the samples 483 were classified as no disease and 77 were classified as having disease samples incorrectly classified are shown in Figure 5.9 Since this study includes imbalanced classes looking at the other metrics as follows and has an AUC value =0.95 depicted in Figure 5.10 of ROC curve.

Accuracy: 0.89

ROC-AUC: 0.95

Classification Table

	Precision	Recall	f1-score	Support
0	0.88	0.98	0.93	3539
1	0.00	0.00	0.00	1378
Accuracy			0.89	4917
Macro Average	0.90	0.81	0.89	4917
Weighted Average	0.52	0.89	0.88	4917

Table 5.8 Classification Table of Gradient Boosting

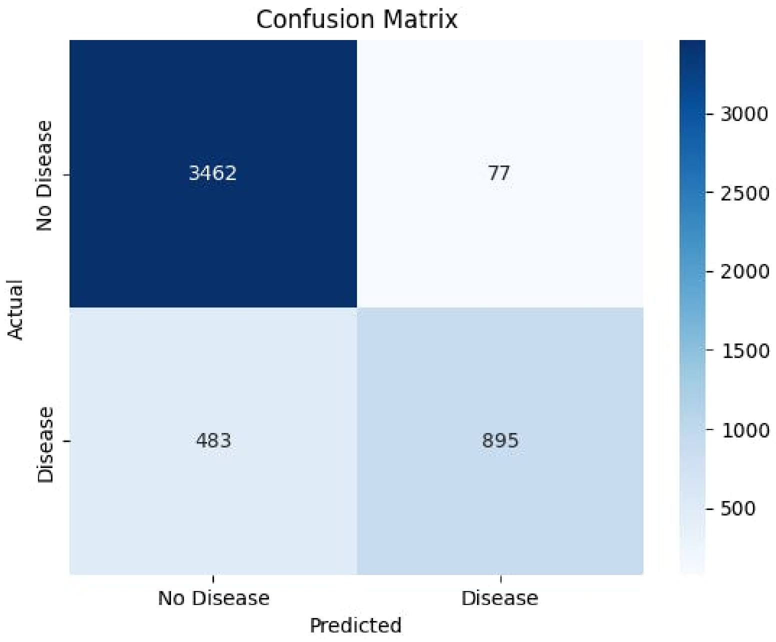


Figure 5.9 Confusion Matrix of Gradient boosting

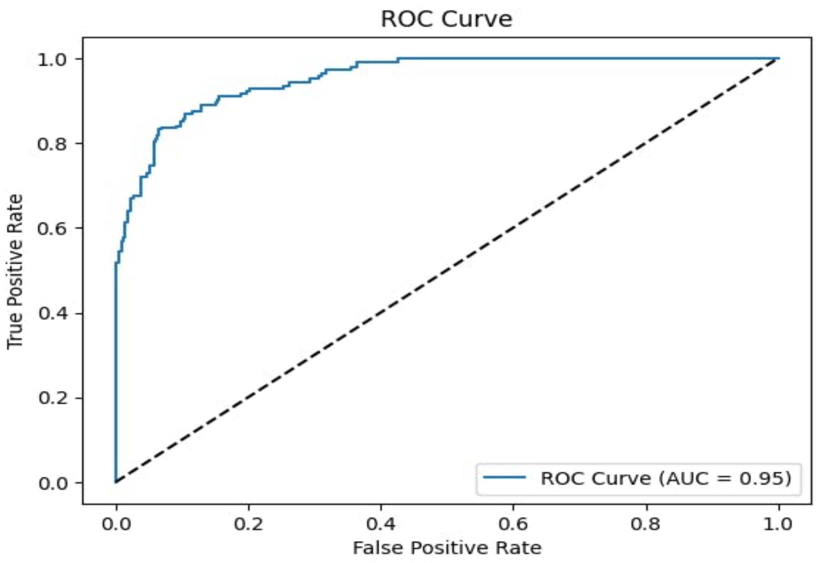


Figure 5.10 ROC curve of Gradient Boosting

5.2.2 Random Forest Model

The Random Forest Model has accuracy of 100 percent in classifying 3536 of no disease cases and 1377 of disease cases. Out of the samples 3 were classified as no disease and 1 were classified as having disease samples incorrectly are shown in Figure 5.11 Since this study includes imbalanced classes looking at the other metrics as follows and has an AUC value =1.00 depicted in Figure 5.12 of ROC curve.

Accuracy: 1.00

ROC-AUC: 1.00

Classification Table

	Precision	Recall	f1-score	Support
0	1.00	1.00	1.00	3539
1	1.00	1.00	1.00	1378
Accuracy			1.00	4917
Macro Average	1.00	1.00	1.00	4917
Weighted Average	1.00	1.00	1.00	4917

Table 5.9 Classification Table of Random Forest

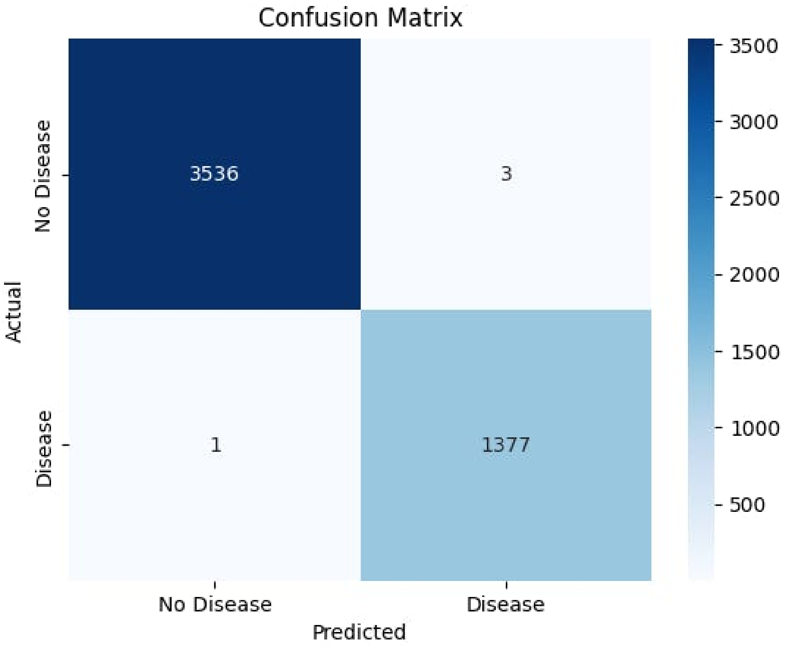


Figure 5.11 Confusion matrix of Random Forest

By comparing the ML models that is Gradient Boosting and Random Forest models, the Random Forest has the highest F1 score of 100 percent, when compared to Gradient Boosting model. From the performance of each model using a Receiver Operating Curve (ROC), the Random Forest has the maximum Area Under the Curve (AUC) value of 1.00.

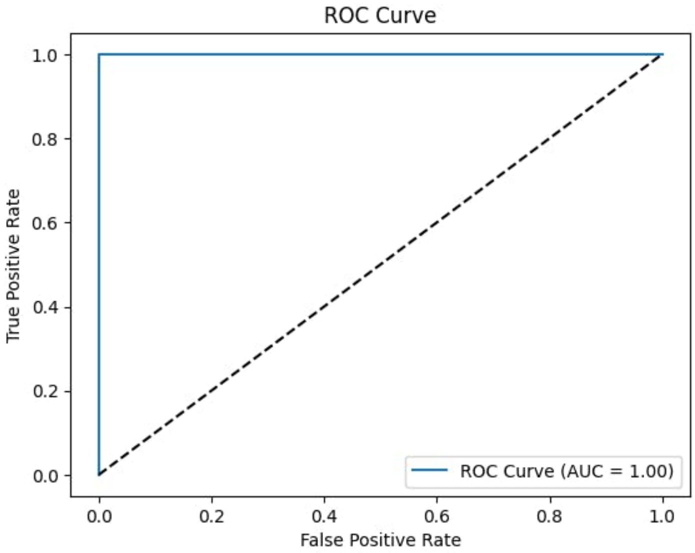


Fig 5.12 ROC Curve

By comparing the ML models that is Gradient Boosting and Random Forest models, the Random Forest has the highest F1 score of 100 percent, when compared to Gradient Boosting model. From the performance of each model using a Receiver Operating Curve (ROC), the Random Forest has the maximum Area Under the Curve (AUC) value of 1.00.

Hence the Random Forest is the best model to classify the dataset into no disease and disease classes. Now identifying the most significant risk feature contributing to liver disease using Random Forest Figure 5.11

Top Contributing Features:

Feature	
Alkphos Alkaline Phosphotase	0.172592
Sgpt Alamine Aminotransferase	0.150853
Sgot Aspartate Aminotransferase	0.148141
Total Bilirubin	0.129187
Total Protiens	0.109684

Table 5.10 Top Contributing Features

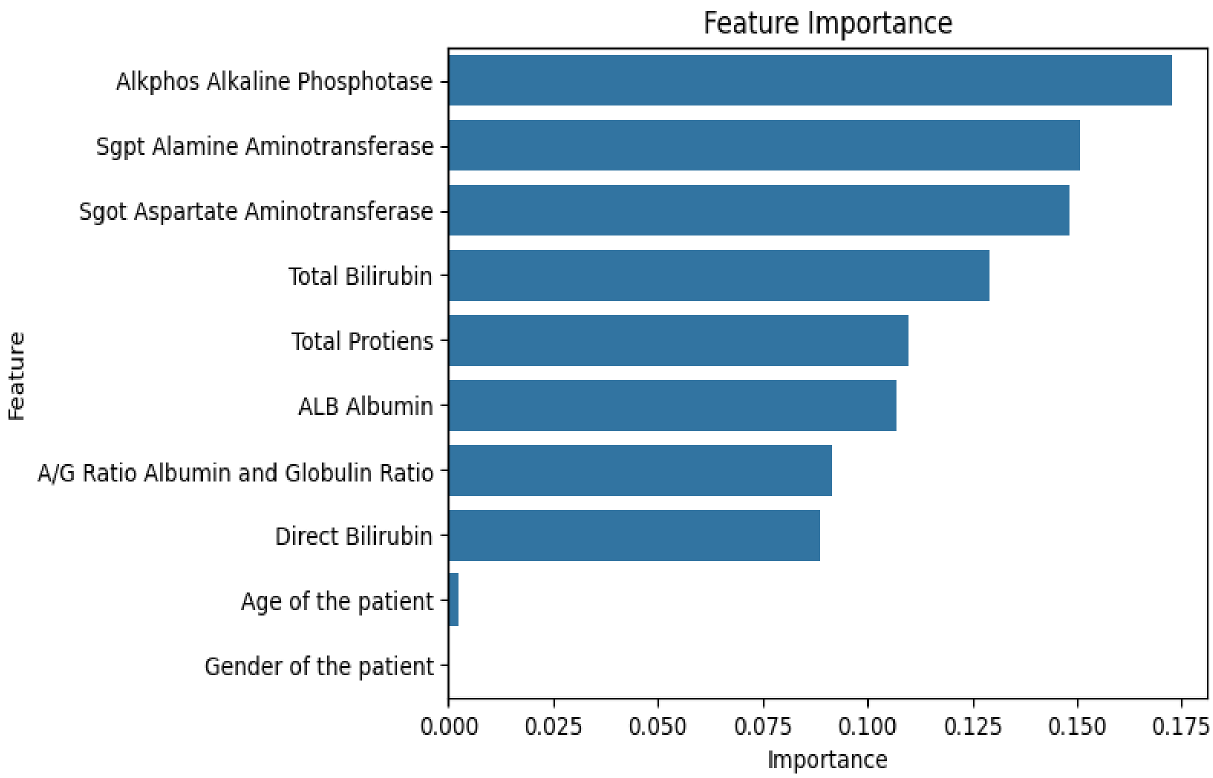


Fig 5.11 Feature Importance

From the Figure 5.11 of feature importance from Random Forest it is evident that Alkphos Alkaline Phosphate has highest rank and thereafter comes Sgpt Alamine Aminotransferase and Sgot Aspartate Aminotransferase. While Alkphos Alkaline Phosphate level plays a major role in the causing liver diseases like hepatitis, cirrhosis or liver cancer.

5.3 COMPARING THE MODELS

Model	Score	Test score
Logistic regression	0.71	0.40
Support Vector Machine	0.72	0.37
Gradient Boosting	0.89	0.94
Random Forest	1.00	0.98

Table 5.11 Comparing the Models

Random forest model has the highest accuracy from the above table it clear that Random Forest is the best predictive model for the prediction of liver disease.

5.4 Building Deep Learning Models

5.4.1 Fully Connected Neural Network (FCNN) Model has accuracy of 97 percent in classifying 2314 of no disease cases and 883 of disease cases. 45 were classified as no disease and 36 were classified as having disease samples wrongly classified. Figure 5.12 showing the training and validation metrics during the training process of a model over multiple epochs:

1. Accuracy Plot (Left):

   The blue line represents the training accuracy.

   The orange line represents the validation accuracy.

2. Loss Plot (Right):

   The blue line represents the training loss.

   The orange line represents the validation loss.

FCNN model executed with 50 training epochs

Every neuron in one layer of a fully connected neural network (FCNN), a deep learning model, is connected to every other neuron in the layer above. According to the statement, the FCNN was trained for 50 epochs, which implies that in order to identify patterns and modify its weights, the model processed the complete dataset 50 times.

Test Loss: 0.4832

Test Accuracy: 0.7219

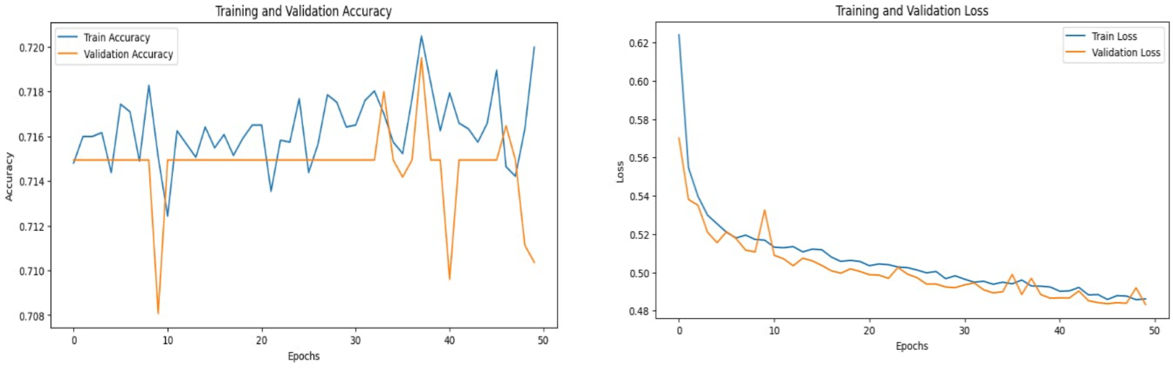


Fig 5.12 Training and Validation graph at Epoch =50

FCNN model executed with 100 training epochs

Test Loss: 0.14798225462436676

Test Accuracy: 0.9752898216247559

Test AUC: 0.9943071007728577

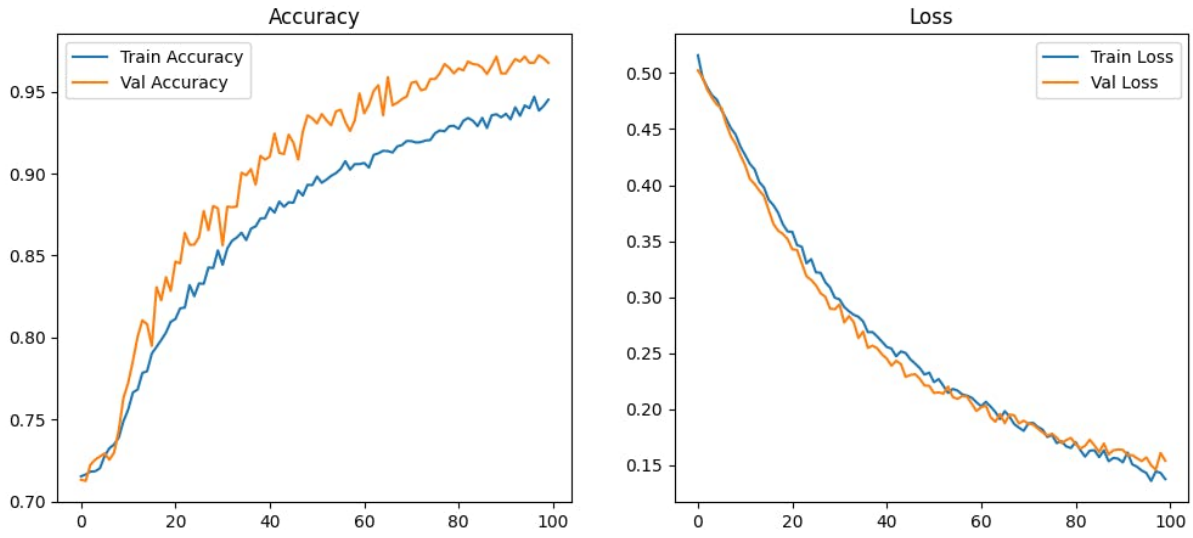


Figure 5.12 Training and Validation graph at Epoch =100

Confusion Matrix:

```
[[2314   36]
 [ 45   883]]
```

Classification Report:

	Precision	Recall	f1-score	Support
0	0.98	0.98	0.98	2350
1	0.96	0.95	0.96	928
Accuracy			0.98	3278
Macro Average	0.97	0.97	0.97	3278
Weighted Average	0.98	0.98	0.98	3278

Table 5.12 Classification Report of FCNN

The training and validation metrics indicate that the model has achieved strong performance and effective generalization. The training and validation accuracy steadily increased over the epochs, stabilizing around 95%, while the corresponding loss values steadily decreased and converged to a low value of approximately 0.1. The close alignment between the training and validation curves demonstrates that the model has effectively learned patterns in the data without overfitting. Additionally, the validation accuracy occasionally surpasses the training accuracy, suggesting the use of regularization techniques to enhance generalization.

5.4.2 Wide and Deep Neural Network

The Wide and Deep Neural Network model has accuracy of 97 percent in classifying 2296 of no disease cases and 894 of disease cases. 34 were classified as no disease and 54 were classified as having disease samples wrongly classified. Figure 5.13 showing the training and validation metrics during the training process of a model over multiple epochs.

WDNN model executed with 50 training epochs

Test Loss: 0.4924

Test Accuracy: 0.7272

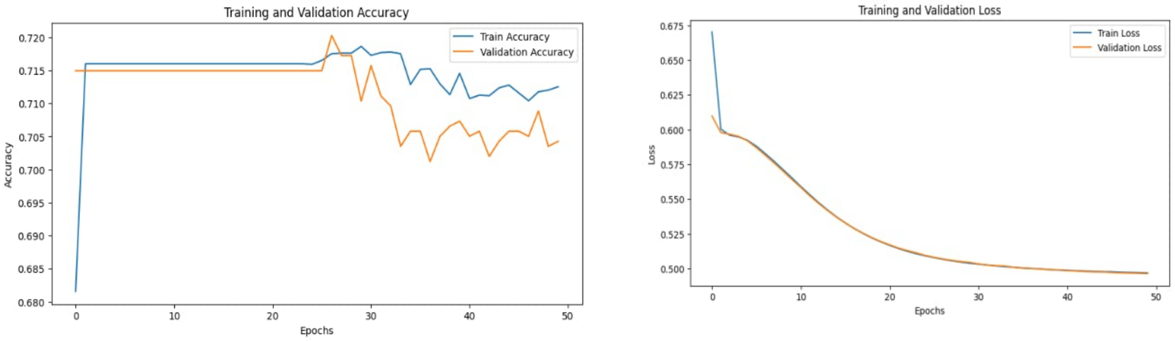


Figure 5.13 Training and Validation graph at Epoch =50

WDNN model executed with 100 training epochs

Test Loss: 0.14817795157432556

Test Accuracy: 0.9731543660163879

Test AUC: 0.9937125444412231

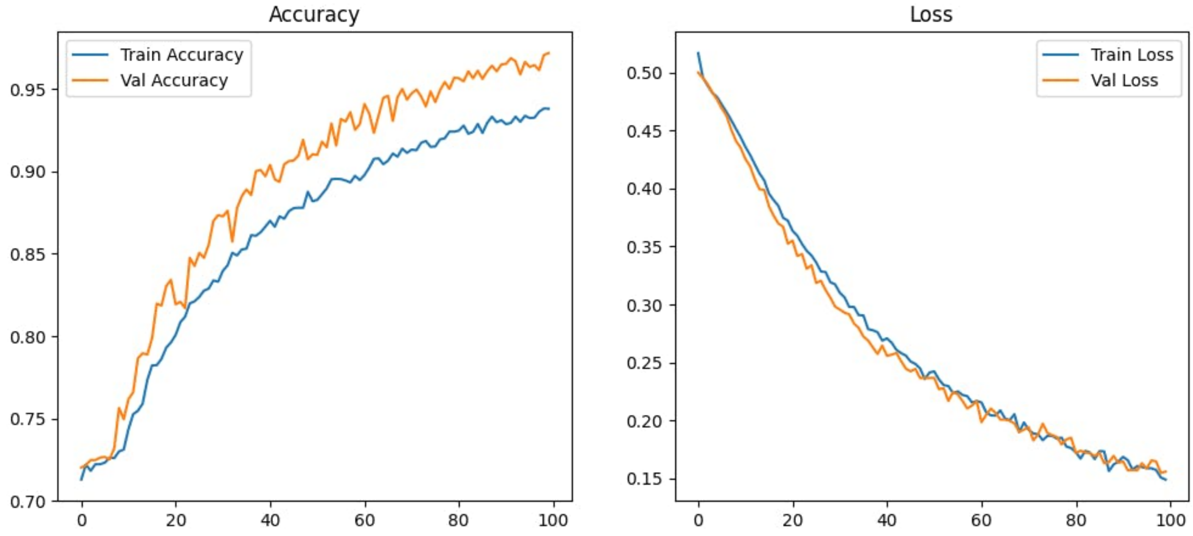


Figure 5.14 Training and Validation graph at Epoch = 100

Classification Report

	Precision	Recall	f1-score	Support
0	0.99	0.98	0.98	2350
1	0.94	0.96	0.95	928
Accuracy			0.98	3278
Macro Average	0.96	0.97	0.97	3278
Weighted Average	0.97	0.97	0.97	3278

Table 5.13 Classification Report of WDNN

```
Confusion Matrix:
[[2296   54]
[ 34   894]]
```

The training and validation accuracy plots show a steady improvement, with training accuracy reaching ~96% and validation accuracy slightly surpassing it at ~97%, indicating effective learning and good generalization. Similarly, both training and validation losses decrease consistently, demonstrating model convergence and the absence of overfitting. The close alignment between training and validation metrics suggests strong generalization capability, possibly enhanced by regularization techniques. These results highlight the model's high predictive performance.

CONCLUSION

In this study, multiple machine learning and deep learning techniques were applied to analyse liver disease data and identify key contributing factors. Initially, Principal Component Analysis (PCA) and K-Means clustering were implemented for dimensionality reduction and pattern identification, followed by evaluation using Logistic Regression and Support Vector Machine (SVM) models. In the second phase, data was assessed using Gradient Boosting and Random Forest, where Random Forest achieved the highest accuracy and was further utilized to determine the most significant features contributing to liver disease. Finally, deep learning approaches, including a Fully Connected Neural Network (FCNN) and a Wide & Deep Neural Network (WDNN), were employed, with both models yielding comparable test accuracy and loss. The findings highlight the effectiveness of Random Forest in feature selection and demonstrate the potential of deep learning for predictive analysis in liver disease classification.

REFERENCES

1. Abdalrada, A. S., Yahya, O. H., Alaidi, A. H. M., Hussein, N. A., Alrikabi, H. T., & Al-Quraishi, T. A. Q. (2019). A predictive model for liver disease progression based on logistic regression algorithm. *Periodicals of Engineering and Natural Sciences (PEN)*, *7*(3), 1255-1264.

2. Al Telaq, B. H., & Hewahi, N. (2021, October). Prediction of liver disease using machine learning models with PCA. In *2021 International Conference on Data Analytics for Business and Industry (ICDABI)* (pp. 250-254). IEEE.

3. Bhusnurmath, R.A., Betageri, S. (2024). Performance Comparison of Machine Learning and Deep Learning Algorithms for Liver Disease Detection. In: Shetty, N.R., Prasad, N.H., Nalini, N. (eds) Advances in Computing and Information. ERCICA 2023. Lecture Notes in Electrical Engineering, vol 1104. Springer, Singapore.

4. Chen, L., Xu, M., Huang, Q., Liu, Y., & Ren, W. (2022). Clinical significance of albumin to globulin ratio among patients with stroke-associated pneumonia. *Frontiers in Nutrition*, *9*, 970573.

5. Ganie, S. M., & Pramanik, P. K. D. (2024). A comparative analysis of boosting algorithms for chronic liver disease prediction. *Healthcare Analytics*, *5*, 100313.

6. Gupta, K., Jiwani, N., Afreen, N., & Divyarani, D. (2022, April). Liver disease prediction using machine learning classification techniques. In *2022 IEEE 11th International conference on communication systems and network technologies (CSNT)* (pp. 221-226). IEEE.

7. Md, A. Q., Kulkarni, S., Joshua, C. J., Vaichole, T., Mohan, S., & Iwendi, C. (2023). Enhanced preprocessing approach using ensemble machine learning algorithms for detecting liver disease. *Biomedicines*, *11*(2), 581.

8. Mostafa, F., Hasan, E., Williamson, M., & Khan, H. (2021). Statistical machine learning approaches to liver disease prediction. *Livers*, *1*(4), 294-312.

9. Sompura, D. U., Tripathy, B. K., Tripathy, A., & Kasat, I. R. (2024, January). Comparative Analysis of Deep Learning-Based Hybrid Algorithms for Liver Disease Prediction. In *International Conference on Advances in Distributed Computing and Machine Learning* (pp. 1-13). Singapore: Springer Nature Singapore.

10. Zhao, J., Wang, P., & Pan, Y. (2022). Predicting liver disorder based on machine learning models. *The Journal of Engineering*, *2022*(10), 978-984.