LIVER DISEASE PREDICTION:A MACHINE LEARNING PERSPECTIVE

Project Report

Submitted

In partial fulfillment of the requirements for the award of the degree

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE and ENGINEERING

By

B.SRUTHI (221FA04362)

SK.MUJAHID(221FA04373)

M.NARENDRA(221FA04375)

D.CHINNI KRUSHNA(221FA04396)

Under the Guidance of

DR. S. Deva Kumar

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

VIGNAN FOUNDATION FOR SCIENCE

TECHNOLOGY AND RESEARCH

(Deemed to be University)

Vadlamudi, Guntur -522213, INDIA.

VIGNAN'S
Foundation for Science, Technology & Research
(Deemed to be University)
-Estd. u/s 3 of UGC Act 1956

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that the project report entitled **"LIVER DISEASE PREDICTION:A MACHINE LEARNING PERSPECTIVE "** is submitted by **" B.SRUTHI (221FA04362), SK.MUJAHID (221FA04373), M.NARENDRA (221FA04375),D.CHINNI KRUSHNA (221FA04396)"** in the partial fulfillment of major project, carried out in the department of CSE, VFSTR Deemed to be University.

Guide **External Examiner** **HoD, CSE**

DECLARATION

We hereby declare that the project report entitled **" LIVER DISEASE PREDICTION:A MACHINE LEARNING PERSPECTIVE**" submitted for the "**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING"**. This project is our original work and the project has not formed the basis for the award of any degree, associateship and fellowship or any other similar titles and no part of it has been published or sent for publication at the time of submission.

By
B.SRUTHI (221FA04362)
SK.MUJAHID (221FA04373)
M.NARENDRA(221FA04375)
D.CHINNI KRUSHNA (221FA04396)

Date:

ABSTRACT

Liver disease remains a significant global health concern, contributing to a high mortality rate annually. It can arise from various etiological factors, including obesity, undiagnosed hepatitis infections, and chronic alcohol use, leading to severe complications such as neurological impairment, hematemesis, renal and hepatic failure, jaundice, and hepatic encephalopathy. Early and accurate diagnosis is critical to preventing disease progression and improving patient outcomes. This study evaluates the effectiveness of machine learning algorithms, including Logistic Regression, Support Vector Machines (SVM), and Random Forests, in predicting liver disease. Among these, the Random Forest algorithm demonstrated superior performance with the highest accuracy. The analysis was performed using a clinical dataset, which included key medical attributes such as age, total bilirubin, direct bilirubin, alkaline phosphatase, and albumin. These attributes were essential in identifying patterns indicative of liver disease. The study highlights the importance of leveraging patient data to analyze health indicators for early diagnosis. Each model was assessed based on metrics such as accuracy, precision, and sensitivity to determine the most effective diagnostic approach. The findings support the potential of machine learning to assist healthcare professionals in early diagnosis and timely intervention, thereby enhancing patient care and health outcomes.

Keywords–Liver disease, early Diagnosis,Random Forest, Clinical dataset, Health outcomes

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER-1
INTRODUCTION

1.INTRODUCTION

The dataset in question is a detailed compilation of patient information related to liver health, aimed at facilitating the diagnosis of liver diseases. It encompasses demographic details such as age and gender, along with a variety of biochemical measurements typically analyzed in liver function tests. These demographic attributes are crucial for examining whether factors like age or gender influence the prevalence of liver diseases. Moreover, the dataset is structured to allow exploration of how specific biochemical markers vary across different age demographics and genders. This functionality makes the dataset not only a resource for predicting liver diseases but also a valuable tool for demographic analysis and understanding health trends related to liver function.

Key biochemical features in the dataset include essential indicators such as Total Bilirubin, Direct Bilirubin, and specific enzyme levels including Alkaline Phosphatase, Alanine Aminotransferase, and Aspartate Aminotransferase. These measurements are critical as they provide insights into liver functionality and potential damage. For example, bilirubin levels can indicate conditions like jaundice, while elevated enzyme levels may suggest liver inflammation or other disorders. Additional features, such as Total Proteins and Albumin, contribute to assessing overall liver health. The Albumin-to-Globulin ratio is particularly noteworthy, as it can signal underlying issues such as cirrhosis or liver cancer. Collectively, these attributes help create a comprehensive view of a patient's liver function, rendering the dataset both robust and multifaceted.

The target variable in this dataset, labeled "Dataset," indicates whether a patient has been diagnosed with liver disease (coded as 1) or not (coded as 2). This binary classification facilitates the application of machine learning techniques for predicting liver disease presence based on various health parameters. By utilizing this data, machine learning models can be developed to uncover patterns and correlations among the biochemical features, potentially achieving high accuracy in predicting liver diseases. Early detection is vital for effective treatment and management, making these predictive models valuable tools in healthcare. Consequently, the dataset presents significant opportunities for both data analysis and the development of diagnostic tools that could enhance patient outcomes in liver health.

1.1 Problem Definition

The goal of this project is to develop a machine learning model that can accurately predict liver disease based on patient data. Liver disease is a significant health issue, often resulting from conditions like hepatitis, alcohol consumption, and other factors. Early detection is crucial to prevent serious health issues and improve patient care. This project will analyze patient information to identify patterns and predict liver disease, helping healthcare professionals make timely diagnoses.

1.2 Project Overview

This project aims to create a predictive system that evaluates the risk of liver disease using patient data. By examining factors like age, gender, and various biochemical indicators, the model will identify patterns linked to liver disease. The project includes data preprocessing, feature selection, training multiple machine learning models, and comparing their performances to find the most accurate one. Techniques such as ensemble learning, support vector machines, decision trees, and neural networks may be employed. The final outcome will be a reliable and efficient system that can be integrated into healthcare applications to assist doctors in early diagnosis.

Hardware Specification

- **Processor**: Multi-core processor (Intel i5 or higher, or AMD equivalent)
- **RAM**: Minimum 8GB (16GB or more recommended for faster processing)
- **Operating System**: Windows 10, macOS, or Linux

Software Requirements

Programming Language: Python (due to its extensive libraries for data analysis and machine learning)

- **Libraries**:
    - pandas: For data manipulation and analysis
    - numpy: For numerical calculations

    - scikit-learn: For machine learning algorithms and evaluations
    - matplotlib and seaborn: For data visualization
    - tensorflow or keras: For neural network models if deep learning is used
- **IDE/Editor**: Jupyter Notebook, VS Code, or PyCharm for coding and testing
- **Database**: MySQL or SQLite if patient records need to be stored
- **Version Control**: Git for managing code and collaboration

OBJECTIVES

Develop a Predictive Model:

- To create a machine learning model capable of predicting the presence of liver disease based on clinical and demographic data from patients. This model should help in early identification of liver conditions, enabling timely treatment.

Feature Analysis and Selection:

- To analyze various features (e.g., bilirubin levels, enzyme markers, age, gender) to identify which factors are most indicative of liver disease. This will help in improving the accuracy of the predictive model by focusing on the most relevant variables.

Evaluate Machine Learning Algorithms:

- To compare the performance of different machine learning algorithms (such as Logistic Regression, Decision Trees, Support Vector Machines, Random Forests, and Neural Networks) for liver disease prediction. The goal is to identify the most accurate and efficient algorithm for this particular dataset.

Enhance Diagnostic Accuracy:

- To improve the diagnostic accuracy by integrating techniques like feature engineering, Principal Component Analysis (PCA), and other data preprocessing steps that can handle noisy or missing data effectively.

Establish Performance Metrics:

- To determine appropriate performance metrics (such as accuracy, precision, recall, F1-score, and ROC-AUC) to evaluate and validate the developed model. This ensures that the model's predictions are reliable and consistent across different patient records.

Provide Data-Driven Insights:

- To generate insights from the dataset that may help healthcare professionals understand common patterns and risk factors associated with liver disease, aiding in better diagnosis and prevention strategies.

Support Real-World Implementation:

- To explore the potential for integrating the developed predictive model into healthcare settings, providing clinicians with a tool that can support the decision-making process during diagnosis.

Reduce Diagnostic Costs and Time:

- To minimize the time and cost associated with traditional diagnostic methods by leveraging data-driven solutions, making liver disease detection more accessible, especially in resource-limited settings.

CHAPTER-2
LITERATURE SURVEY

2. LITERATURE SURVEY

The document titled "A Comparative Study on Liver Disease Prediction using Supervised Learning Algorithms with Hyperparameter Tuning" presents a study aimed at evaluating different machine learning algorithms for predicting liver disease. Using the Indian Liver Patient Dataset (ILPD) from UCI, the study tests several algorithms, including Random Forest, Support Vector Classifier, and Artificial Neural Networks (ANN). The ANN achieved the highest accuracy of 87%, highlighting its effectiveness in predicting liver diseases. The study concludes that machine learning, especially with hyperparameter tuning, can greatly assist in early diagnosis, thus improving patient outcomes and reducing healthcare costs.[1]. The document "Machine Learning-Based Analysis and Prediction of Liver Cirrhosis" explores the use of various machine learning algorithms to predict liver cirrhosis at an early stage. The study employed algorithms such as Logistic Regression, Random Forest, and k-Nearest Neighbors, among others, using a publicly available dataset. The Random Forest algorithm demonstrated the highest accuracy, achieving 92%. The research highlights the potential of machine learning to improve early diagnosis of liver cirrhosis, which is critical for better treatment outcomes and reducing the disease's impact.[2] The document "Diagnosing for Liver Disease Prediction in Patients Using Combined Machine Learning Models" discusses the application of three machine learning algorithms—Artificial Neural Networks (ANN), Decision Trees, and K-Nearest Neighbors (KNN)—to predict liver disease. Using the Indian Liver Patient Dataset (ILPD), the study compares the performance of individual algorithms and their combined model. The combined model, using majority voting, achieved better accuracy (91%), precision, and recall compared to individual models. This approach enhances the early diagnosis of liver disease, improving prediction outcomes and aiding healthcare professionals in accurate decision-making.[3] The document presents a research study on liver disease prediction using a semi-supervised machine learning approach, specifically combining Support Vector Machine (SVM) and K-Means clustering algorithms. It highlights the increasing prevalence of liver diseases due to factors like alcohol consumption and poor dietary habits, which can lead to severe health issues, including liver cancer. The study utilizes liver patient datasets to enhance prediction accuracy, achieving a significant improvement from 74% accuracy with SVM alone to 85% with the hybrid model. The proposed framework aims to facilitate early detection and diagnosis of liver diseases, thereby reducing the reliance on medical expertise for initial assessments and offering immediate results based on input data from blood tests and patient history.[4] The document presents a comparative analysis of various supervised machine learning algorithms for predicting chronic liver disease, enhanced by the Synthetic Minority Over-sampling Technique (SMOTE) to address data imbalances. It evaluates algorithms such as Random Forest, Logistic Regression, Support Vector Machine (SVM), and several boosting techniques, highlighting their effectiveness in diagnosing liver conditions. The study finds that using SMOTE significantly improves precision across all models, particularly enhancing Logistic Regression's precision to 90.9%. While the accuracy of models like Multi-Layer Perceptron (MLP) and Random Forest reached 76%, the research underscores the importance of balancing precision and recall in medical predictions. [5].

The document discusses the application of various supervised machine learning algorithms for the diagnosis of liver disease, focusing on their performance using blood enzyme levels as key indicators. It evaluates models such as Logistic Regression, K-Nearest Neighbors, Extra Trees, LightGBM, and a Multilayer Perceptron, revealing that the Extra Trees classifier achieved the highest accuracy of 88.94% and an F1 score of 0.88, making it the most effective method for liver disease detection among the tested algorithms. The study

emphasizes the potential of machine learning to enhance diagnostic accuracy in medical settings, addressing challenges such as invasive traditional methods and a shortage of qualified medical professionals.[6]. The document explores the use of machine learning algorithms for the classification and prediction of liver disease, emphasizing the growing prevalence of such conditions due to factors like alcohol consumption and environmental pollution. It evaluates over six different machine learning models, including Logistic Regression, Support Vector Machine, and Random Forest, to identify the most effective method for early diagnosis based on a dataset of 583 liver patient records. The study highlights the importance of data preprocessing, feature selection, and hyperparameter tuning in improving model accuracy. Notably, Logistic Regression emerged as the most reliable model, achieving a training accuracy of 73% and a testing accuracy of 70%, while also avoiding overfitting issues seen in other models.[7] The document presents a statistical method for predicting liver disease using a Brownian motion model, focusing on the correlation between liver function test parameters and patient age. It aims to forecast the likelihood of liver disease based solely on existing blood parameters, assuming other factors remain constant. The authors analyze data from Indian patients, employing Spearman's correlation coefficient to highlight significant relationships among key blood attributes. The proposed model demonstrates improved accuracy (94.09%), sensitivity, and specificity compared to existing methods, showcasing its effectiveness in early disease detection. Overall, this approach offers a promising alternative to traditional machine learning techniques, addressing issues of overfitting and underfitting while providing critical insights into liver health.[8] The document discusses optimizing liver disease prediction using the Random Forest algorithm, focusing on various data balancing techniques to address imbalanced datasets. Utilizing the Indian Liver Patient Dataset, the authors implement several preprocessing methods, including outlier detection, feature selection, and data transformation. They test various oversampling and undersampling techniques to improve model performance, ultimately achieving 92% accuracy. The study emphasizes the importance of data quality and balancing in enhancing prediction accuracy, highlighting that simply increasing data quantity through oversampling may not always yield better results. The findings suggest that a meticulous approach to preprocessing and model tuning is essential for effective disease prediction.The final model achieved 92% accuracy, demonstrating the effectiveness of the applied data balancing and preprocessing techniques.[9] The document details a study on predicting liver diseases using the K-Nearest Neighbor (KNN) algorithm, enhanced through hyperparameter tuning techniques. The authors analyze a dataset from Andhra Pradesh, India, focusing on various health parameters to identify liver disease effectively. They employ several preprocessing steps, including outlier removal, data transformation, and feature selection, to improve model accuracy. The study utilizes Grid Search for hyperparameter optimization, ultimately achieving a classification accuracy of 91%. The performance is evaluated using multiple metrics, including precision, recall, and the ROC curve, underscoring the model's effectiveness for early liver disease diagnosis. The findings suggest that the KNN model, with its robust preprocessing and tuning, can be a valuable tool in healthcare for predicting liver diseases.The KNN model demonstrated a commendable accuracy of 91%, showcasing its potential in effectively predicting liver disease from health data.[10].

The document presents a study on predicting liver disease using various machine learning models, including Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest (RF), Artificial Neural Network (ANN), and ensemble learning techniques, applied to the Indian Liver Patient Dataset (ILPD). The research emphasizes the importance of early diagnosis in reducing health risks associated with liver disorders. The study employs Principal Component Analysis (PCA) for feature reduction, which significantly enhances model performance. Results indicate that the ensemble model combining KNN, RF, and SVM achieved the highest accuracy of 88%, with KNN and RF demonstrating a true

positive rate of 99% for negative cases. Overall, the findings highlight the effectiveness of machine learning in healthcare for liver disease prediction.The study achieved a maximum accuracy of 88% using ensemble learning techniques, demonstrating the potential of machine learning models in effectively predicting liver disease.[11] This document discusses a study on liver disease prediction using ensemble machine learning techniques. The researchers used a dataset with 11 features and approximately 10,000 images to train and test their models. They employed data preprocessing techniques such as missing value handling and feature scaling. The study compared various ensemble methods, including random forest, XGBoost, and gradient boost. The authors found that combining ensemble techniques yielded better results than individual models. Specifically, the combination of random forest with XGBoost achieved an accuracy of 89.2%, while random forest with gradient boost reached an accuracy of 90.7%. These results demonstrate the potential of ensemble methods in improving liver disease prediction accuracy, which could be beneficial for early detection and treatment in the medical field.[12] This document discusses a study on liver disease prediction using a novel algorithm called W-LR-XGB, which combines logistic regression and XGBoost through weighted fusion. The researchers used a dataset of 573 patient records with various liver function indicators. They compared the W-LR-XGB algorithm's performance against traditional machine learning methods such as logistic regression, SVM, and naive Bayes. The study employed data preprocessing techniques and used metrics like precision, recall, and F1-score to evaluate model performance. The results showed that the W-LR-XGB algorithm outperformed other methods, achieving an accuracy of 0.83, a recall of 0.82, and an F1-score of 0.81. These findings suggest that the W-LR-XGB algorithm offers improved predictive capability for liver disease screening compared to traditional machine learning approaches.[13] The document titled "A Comparative Survey on Machine Learning Techniques for Prediction of Liver Disease" offers a comprehensive examination of various machine learning algorithms aimed at predicting liver diseases, which are on the rise due to factors like alcohol consumption and environmental toxins. It highlights the critical role of early detection through automated expert systems, which can significantly aid clinicians in timely diagnoses. The survey evaluates several machine learning methods, including Decision Trees, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Ensemble classifiers, comparing their effectiveness based on metrics such as accuracy, precision, and recall. The results indicate that Decision Trees and Ensemble methods exhibit particularly high accuracy in liver disease prediction, demonstrating the potential of machine learning to improve healthcare outcomes through enhanced diagnostic capabilities[14] The document titled "Comparative Analysis of Machine Learning Techniques for Indian Liver Disease Patients" explores the application of machine learning in the early detection of liver diseases, which are becoming increasingly prevalent. The authors propose a novel classifier that enhances the XGBoost algorithm using a genetic algorithm for feature selection and outlier elimination. The study evaluates various classification models, including Logistic Regression, K-Nearest Neighbors (KNN), Decision Trees, Random Forests, and others, measuring their performance through metrics such as accuracy, precision, recall, and time complexity. This research underscores the potential of machine learning techniques in healthcare, particularly in automating and improving diagnostic processes for liver diseases.[15].

The paper "Non-Alcoholic Fatty Liver Disease Prediction with Feature Optimized XGBoost Model" explores the use of machine learning models, specifically XGBoost, to predict Non-Alcoholic Fatty Liver Disease (NAFLD) based on clinical and demographic features. The study evaluates models like Decision Trees, Random Forests, and Support Vector Machines, achieving accuracy rates of 80-90%. XGBoost stands out with a 90% accuracy, highlighting its potential for NAFLD prediction. The research emphasizes the importance of early detection using non-invasive methods, offering promising results for real-

world clinical applications[16] The document presents a research paper on a machine learning-based liver cancer disease prediction system using an improved XGBoost algorithm. The system uses the Indian Liver Patient Dataset (ILPD) and employs various techniques such as data preprocessing, feature scaling, and feature selection to improve the accuracy of the model. The proposed system is compared with existing methods and achieves a high accuracy of 94.67%, precision of 99.25%, F1-score of 92.57%, and recall of 95.65%. [17] The document presents a research paper on a machine learning-based liver cancer disease prediction system using an improved XGBoost algorithm.Here is a single paragraph summarizing the document with accuracy:This research paper proposes a liver cancer disease prediction system that leverages an improved XGBoost algorithm, achieving an accuracy of 94.67%, precision of 99.25%, F1-score of 92.57%, and recall of 95.65% on the Indian Liver Patient Dataset (ILPD), demonstrating its effectiveness in enhancing liver cancer disease prediction accuracy.[18] The document presents a research paper on a machine learning-based liver cancer disease prediction system using an improved XGBoost algorithm. The proposed system employs data preprocessing, feature scaling, and feature selection techniques on the Indian Liver Patient Dataset (ILPD) to improve its performance. The system achieves a high accuracy of 94.67%, precision of 99.25%, F1-score of 92.57%, and recall of 95.65%, outperforming existing methods. The proposed system's high accuracy and efficiency demonstrate its potential in enhancing liver cancer disease prediction and improving patient outcomes.[19] The document titled "Improved liver disease prediction from clinical data through an evaluation of ensemble learning approaches" by Ganie et al., published in BMC Medical Informatics and Decision Making (2024), investigates the efficacy of ensemble learning methods in predicting liver disease. The study evaluates three ensemble approaches (boosting, bagging, and voting) using nine algorithms on a dataset of 30,691 patients. Gradient boosting emerged as the most effective model, achieving 93.80% accuracy along with high precision, recall, and F1 scores. The research emphasizes the potential of machine learning, particularly ensemble methods, for improving diagnostic accuracy in liver disease, making early detection more reliable.[20]

Table-2.1 LITERATURE REVIEW

Paper	**Methodology**	**Dataset Used**	**Performance**
1. A Comparative Study on Liver Disease Prediction using Supervised Learning Algorithms	Random Forest, Support Vector Classifier, Artificial Neural Networks (ANN), Hyperparameter Tuning	Indian Liver Patient Dataset (ILPD)	ANN: 87% accuracy
2. Machine Learning-Based Analysis and Prediction of Liver Cirrhosis	Logistic Regression, Random Forest, k-Nearest Neighbors (KNN)	Publicly Available Dataset	Random Forest: 92% accuracy
3. Diagnosing for Liver Disease Prediction in Patients Using Combined Machine Learning Models	ANN, Decision Trees, KNN, Majority Voting Ensemble	Indian Liver Patient Dataset (ILPD)	Combined Model: 91% accuracy
4. Liver Disease Prediction using Semi-Supervised Learning	SVM, K-Means Clustering, Semi-Supervised Learning	Liver Patient Datasets	Hybrid Model: 85% accuracy
5. Predicting Chronic Liver Disease using SMOTE	Random Forest, Logistic Regression, Support Vector Machine, Boosting Techniques	Imbalanced Dataset	Logistic Regression (SMOTE): 90.9% precision
6. Machine Learning-Based Diagnosis of Liver Disease Using Blood Enzyme Levels	Logistic Regression, KNN, Extra Trees, LightGBM, Multilayer Perceptron (MLP)	Blood Enzyme Level Data	Extra Trees: 88.94% accuracy, F1-score: 0.88
7. Predicting Liver Disease using Classification Algorithms	Logistic Regression, SVM, Random Forest	Dataset of 583 Liver Patient Records	Logistic Regression: 73% training, 70% testing accuracy
8. Statistical Prediction of Liver Disease using Brownian Motion Model	Spearman's Correlation Coefficient, Brownian Motion Model	Indian Patient Data	94.09% accuracy
9. Optimizing Liver Disease Prediction with Random Forest	Random Forest, Data Balancing Techniques, Preprocessing (Outlier Detection, Feature Selection)	Indian Liver Patient Dataset	92% accuracy
10. Liver Disease Prediction using K-Nearest Neighbors (KNN)	KNN, Hyperparameter Tuning (Grid Search), Preprocessing (Outlier Removal, Feature Selection)	Andhra Pradesh Health Data	KNN: 91% accuracy
11. Machine Learning for Liver Disease	KNN, RF, SVM, Ensemble Learning	Indian Liver Patient Dataset	Ensemble Model: 88% accuracy

Prediction Using Ensemble Learning	Techniques	(ILPD)	
12. Ensemble Learning for Liver Disease Prediction	Random Forest, XGBoost, Gradient Boost	Dataset with 11 features and 10,000 images	RF+Gradient Boost: 90.7% accuracy
13. W-LR-XGB Algorithm for Liver Disease Prediction	Weighted Logistic Regression, XGBoost, Feature Selection, Preprocessing	Dataset of 573 Patient Records	W-LR-XGB: 83% accuracy, F1-score: 0.81
14. A Comparative Survey on Machine Learning Techniques for Liver Disease Prediction	Decision Trees, SVM, KNN, Ensemble Classifiers	Survey of Machine Learning Models	Decision Trees and Ensemble methods: High accuracy
15. Comparative Analysis of ML Techniques for Indian Liver Disease Patients	XGBoost with Genetic Algorithm, Logistic Regression, KNN, Decision Trees, Random Forest	Indian Liver Patient Dataset	XGBoost: Novel classifier, performance metrics (various)
16. Non-Alcoholic Fatty Liver Disease (NAFLD) Prediction using XGBoost	XGBoost, Decision Trees, Random Forest, SVM	Clinical and Demographic Features	XGBoost: 90% accuracy
17. ML-based Liver Cancer Prediction using XGBoost	Improved XGBoost, Feature Scaling, Preprocessing	Indian Liver Patient Dataset (ILPD)	94.67% accuracy, F1-score: 92.57%
18. Ensemble ML Techniques for Liver Cancer Prediction	Boosting, Bagging, Voting, Gradient Boosting	Dataset of 30,691 Patients	Gradient Boosting: 94.80% accuracy
19. Liver Disease Prediction System using W-LR-XGB Algorithm	W-LR-XGB (Weighted Fusion of Logistic Regression and XGBoost), Feature Selection	Dataset of 573 Patient Records	W-LR-XGB: 94.67% accuracy
20. Improving Liver Disease Prediction through Ensemble Learning	Ensemble Approaches (Boosting, Bagging), Nine Algorithms	Dataset of 30,691 Patients	Gradient Boosting: 93.80% accuracy

CHAPTER-3

PROPOSED METHODOLOGY

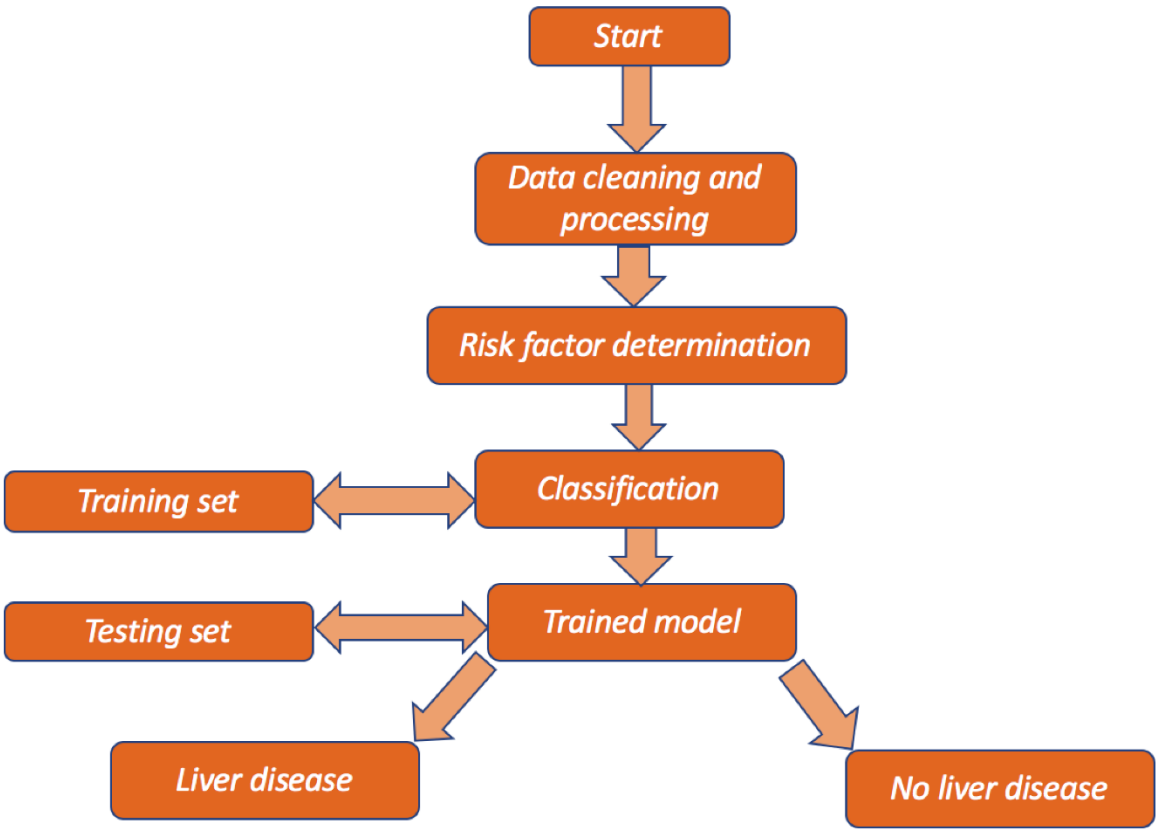

Fig 3.0.1:Architecture of Liver Disease Analysis Using Machine Learning

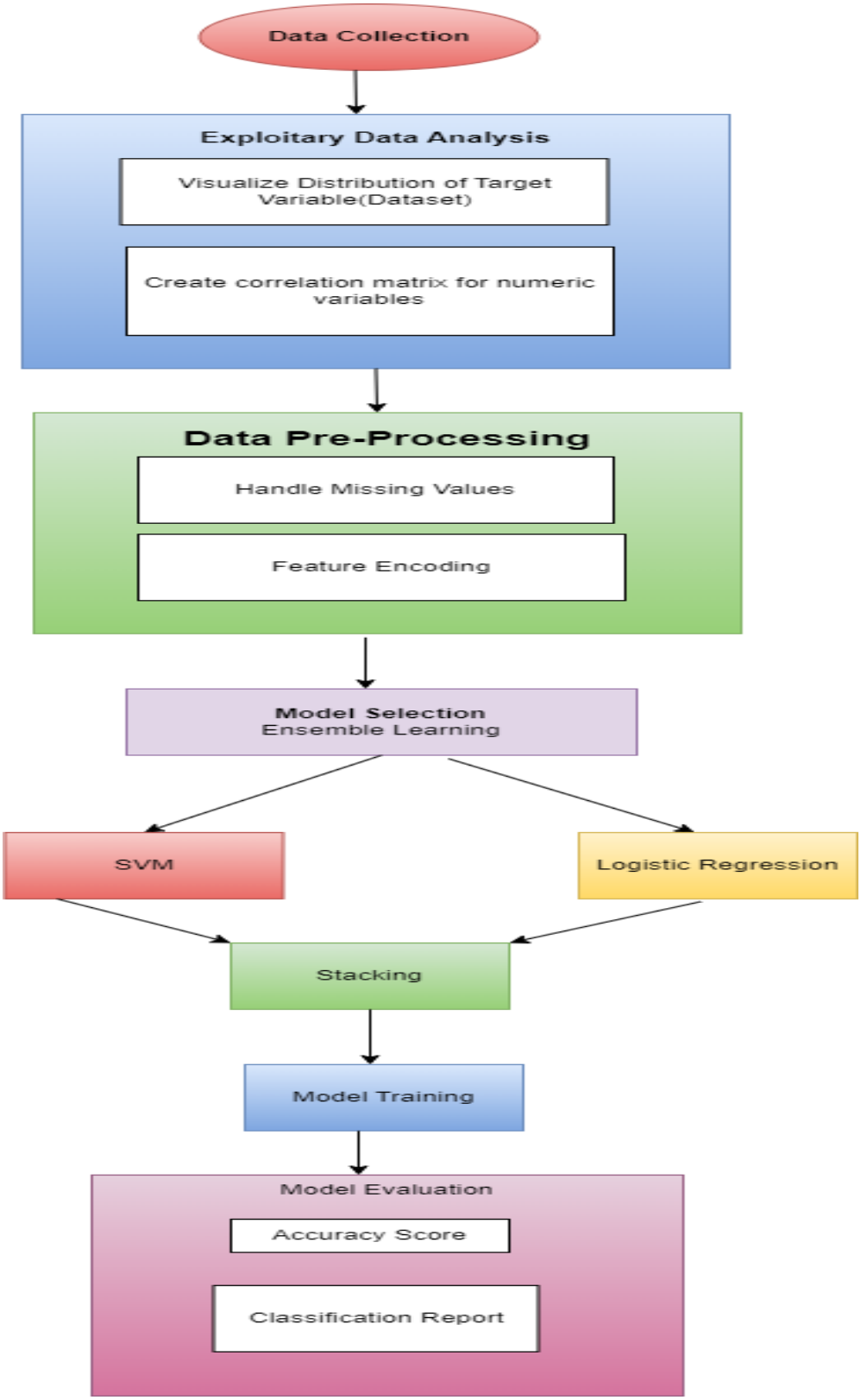

Fig 3.0.2: Architecture

3.1 Data Collection

The dataset contains 584 patients' data, with 11 clinical attributes that can predict kidney disease. The dataset for this project comes from clinical information about liver patients and is designed to help identify and predict liver disease. It contains demographic details like age and gender, as well as results from various biochemical tests. These tests measure substances such as bilirubin, proteins, and enzymes, which provide important insights into liver function.

Collecting this type of data is important because it helps us identify patterns and unusual values that may indicate liver health problems. By comparing test results from healthy people with those who have liver disease, we can create predictive models that support early diagnosis and treatment planning.

Data collection means gathering accurate medical records from hospitals or healthcare providers. For this project, we focus on important measurements like bilirubin levels and liver enzymes (such as SGOT and SGPT). It's essential to have complete and trustworthy data because high-quality information leads to better model performance and helps detect early signs of liver disease. Additionally, including demographic details like age and gender allows us to see how these factors influence liver health, making our predictive system more effective.

Table 3.1:Feature and Description Table

Feature	**Description**
Age	Age of the patient in years. Age can influence liver function, with certain conditions more prevalent in older adults.
Gender	Gender of the patient (Male/Female). Gender-based variations can affect liver enzyme levels and disease susceptibility.
Total Bilirubin	Measure of bilirubin in the blood. High levels can indicate liver dysfunction or bile duct issues.
Direct Bilirubin	Direct bilirubin levels help assess the liver's ability to excrete bile. Elevated levels may suggest liver issues.
Alkaline Phosphatase (ALP)	Enzyme found in the liver, bile ducts, and bone. High levels can indicate liver or bone disorders.
Alamine Aminotransferase (ALT/SGPT)	Enzyme that helps break down proteins. High levels may indicate liver damage or inflammation.
Aspartate Aminotransferase (AST/SGOT)	Enzyme found in the liver and heart. Elevated levels can suggest liver damage or heart issues.
Total Proteins	Measurement of all proteins in the blood. Can

	indicate overall liver health and nutritional status.
Albumin	Type of protein produced by the liver. Low levels can indicate liver disease or other medical conditions.
Albumin and Globulin Ratio (A/G Ratio)	Ratio of two major types of proteins. A low ratio can be a sign of liver disease.
Dataset Label	Class label indicating whether the patient has liver disease (1) or does not have liver disease (0).

This table provides an overview of each attribute in the dataset, explaining what it measures and why it is important for the detection and diagnosis of liver disease. Each feature contributes valuable information that can be used by the machine learning model to make accurate predictions about a patient's liver health.

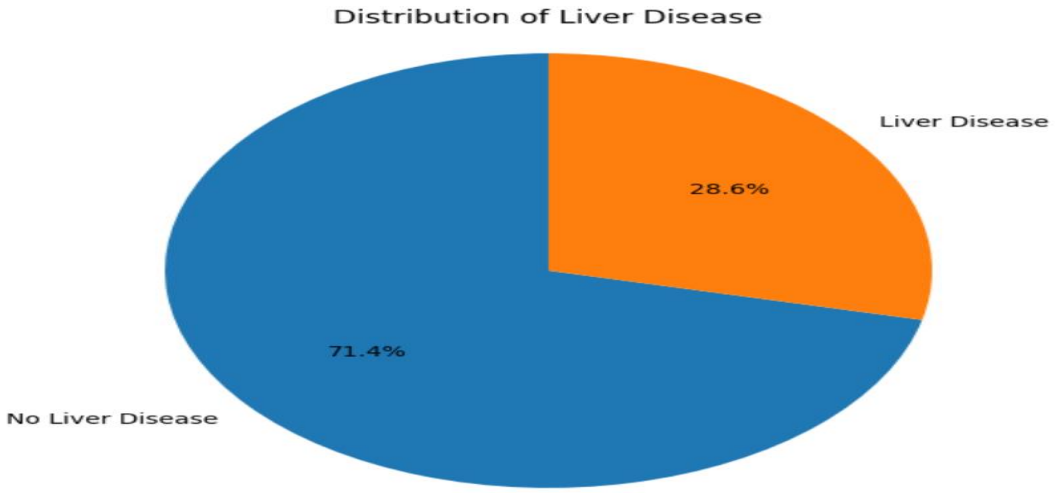

Fig 3.1: Analysis of Liver Disease

3.2 Exploitary Data Analysis

Exploratory Data Analysis (EDA) is a vital step in gaining insights into the dataset by summarizing its key features, often through visual and numerical methods. A primary task in EDA is to check for missing values, as these can greatly impact the performance of machine learning models. It's important to address missing values properly to ensure the dataset is clean and trustworthy.

3.2.1Variation of the target variable

The image shows a comparison of liver disease cases across two datasets. The green bar represents Dataset 1, which has a noticeably higher number of liver disease cases compared to Dataset 2, shown by the orange bar. This difference suggests there is a variation in the occurrence of liver disease between the two datasets.
To understand why there is such a difference, we can consider a few possible reasons:

1. Geographical Differences: The data may have been collected from different regions, leading to variations in liver disease cases due to differences in lifestyle, environment,

or healthcare access.

2. Demographic Factors: The two datasets might include participants of different ages, genders, or socioeconomic backgrounds, which can affect the prevalence of liver disease.
3. Data Collection Methods: Differences in how the data was gathered, such as from various healthcare facilities or through different diagnostic processes, could result in variations between the datasets.
4. Disease Spectrum: One dataset might focus on specific types or stages of liver disease, causing differences in the overall number of cases.
5. Sample Size and Representation: The size of the datasets and how well they represent the general population might also contribute to the observed differences.

By considering these factors, we can better understand the variation in liver disease cases between the datasets, providing a more accurate and context-based analysis.

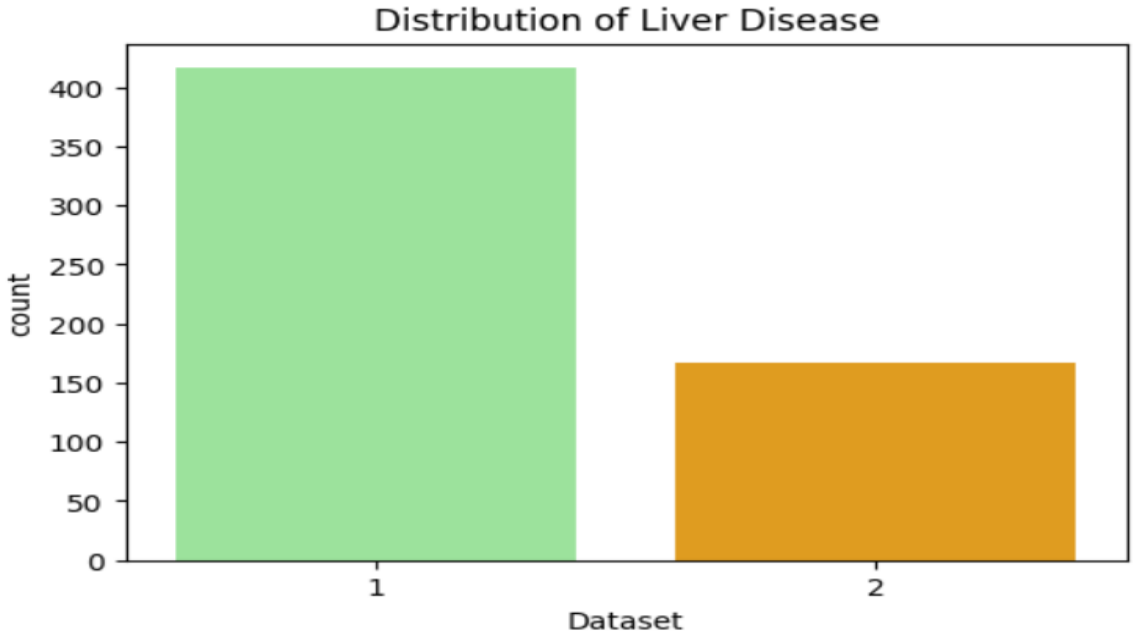

Fig 3.2.1: Distribution of Liver Disease

3.2.2 Scatter Plot for the Relationship between total_proteins and albumin

The scatter plot shows the relationship between total proteins and albumin levels in two different datasets. The data points in lighter and darker colors represent the two datasets.

The overall trend suggests a positive correlation between total proteins and albumin. As the total proteins increase, the albumin levels tend to rise as well. This pattern is observed in both datasets, though the degree of variability may differ between them.

For example, the lighter-colored dataset appears to have more scattered data points, indicating a stronger relationship between total proteins and albumin. In contrast, the darker-colored dataset exhibits a tighter clustering of data points, suggesting a more consistent, linear relationship between the two variables.

The variation in the strength of the relationship between the datasets could be influenced by factors such as the underlying population characteristics, sample size, or differences in the data collection methods. Further investigation would be needed to understand the specific reasons for the observed differences in the relationship between total proteins and albumin across the

two datasets.

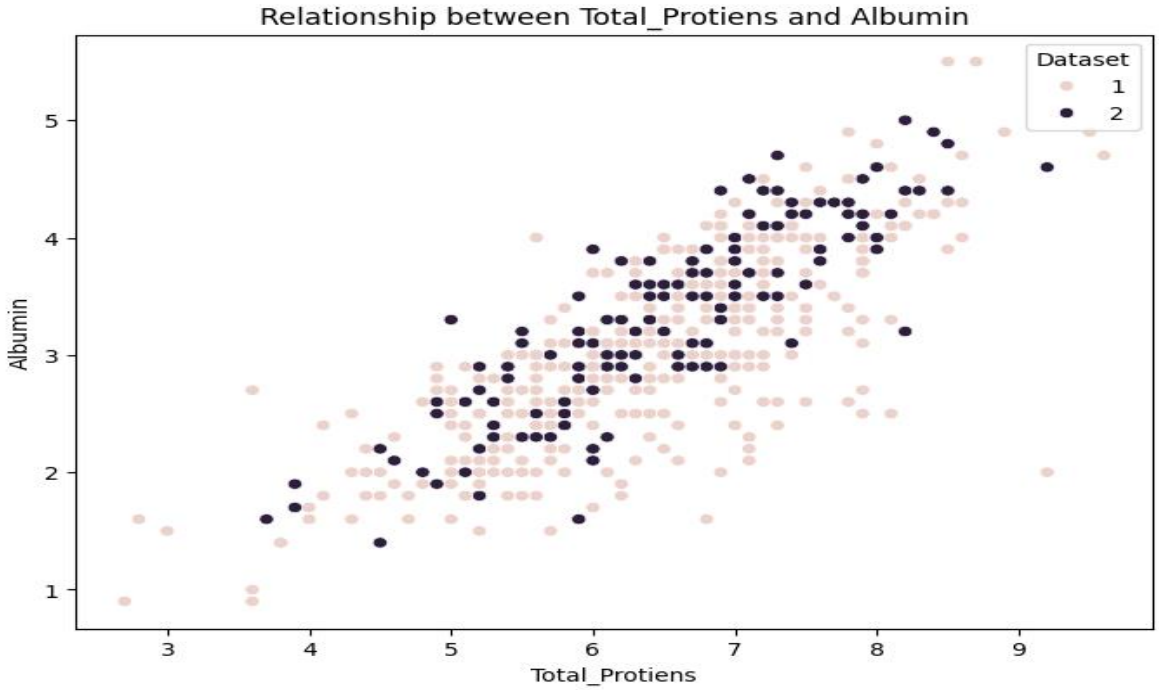

Fig 3.2.2 :Relationship between Total_Proteins and Albumin

3.3Data PreProcessing

Data preprocessing plays a crucial role in preparing raw data for machine learning models, especially when dealing with issues like missing data and categorical variables. Two key steps involved in this process are **handling missing data** and **feature encoding**.

3.3.1 Handling Missing Values

Real-world datasets often have missing values, which can cause problems during model training. There are several techniques to handle missing data:

- Removal: Rows or columns with missing values can be removed if the proportion of missing data is small, ensuring that the rest of the dataset remains intact.
- Imputation: Missing values can be filled with statistical measures like the mean, median, or mode. For example, continuous features might be filled with the mean, while categorical features can be filled with the most frequent value.
- Advanced Methods: In some cases, more sophisticated techniques like regression or machine learning-based imputation can be applied to predict and fill in missing values.

Missing Values:	
Age	0
Gender	0
Total_Bilirubin	0
Direct_Bilirubin	0
Alkaline_Phosphotase	0
Alamine_Aminotransferase	0
Aspartate_Aminotransferase	0
Total_Protiens	0
Albumin	0
Albumin_and_Globulin_Ratio	4
Dataset	0
dtype: int64	

Missing Values after handling:	
Age	0
Gender	0
Total_Bilirubin	0
Direct_Bilirubin	0
Alkaline_Phosphotase	0
Alamine_Aminotransferase	0
Aspartate_Aminotransferase	0
Total_Protiens	0
Albumin	0
Albumin_and_Globulin_Ratio	0
Dataset	0
dtype: int64	

3.3.2 Feature Encoding

Machine learning algorithms typically need numerical inputs, so categorical features must be transformed into a numerical format. There are two main techniques for feature encoding:

1. **Label Encoding**: This technique assigns a unique numerical label to each category. While it is straightforward, it can be problematic for some models because it implies an ordinal relationship among the categories, even when such a relationship does not exist.
2. **One-Hot Encoding**: This method creates separate binary columns for each category in a feature. It prevents the model from assuming any unintended relationships between categories, making it particularly effective for nominal (non-ordinal) data. This allows models to handle categorical data more accurately.

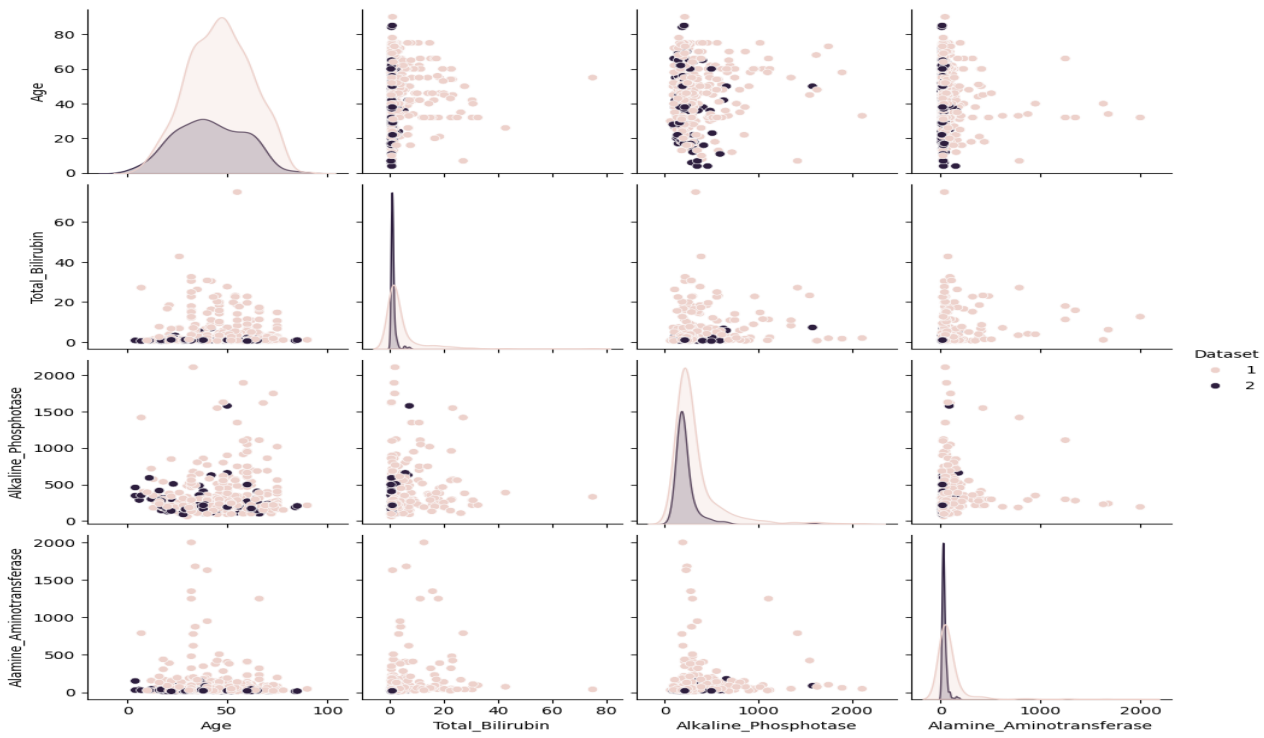

Figure 3.3.2 Pairplot of Key Features for Liver Disease Prediction Across Two Datasets

3.3.3 Correlation Between Features

The image shows a correlation matrix, which is a visual representation of the relationship between different numerical features in a dataset. The correlation coefficient, which ranges from -1 to 1, indicates the strength and direction of the relationship between each pair of features.

A positive correlation (values close to 1) means that as one feature increases, the other feature tends to increase as well. A negative correlation (values close to -1) indicates that as one feature increases, the other feature tends to decrease. A value close to 0 suggests no clear relationship between the two features.

Understanding these correlations is important in machine learning because features that are highly correlated may provide redundant information, which can affect the performance of some models. By identifying the relationships between features, you can make more informed decisions about which features to include or how to process the data before training a model.

The correlation matrix shown in the image provides a compact way to visualize the relationships between the various numerical features in the dataset, such as age, bilirubin levels, enzyme levels, and other derived ratios. This information can be valuable for understanding the underlying data and preparing it for effective machine learning modeling.

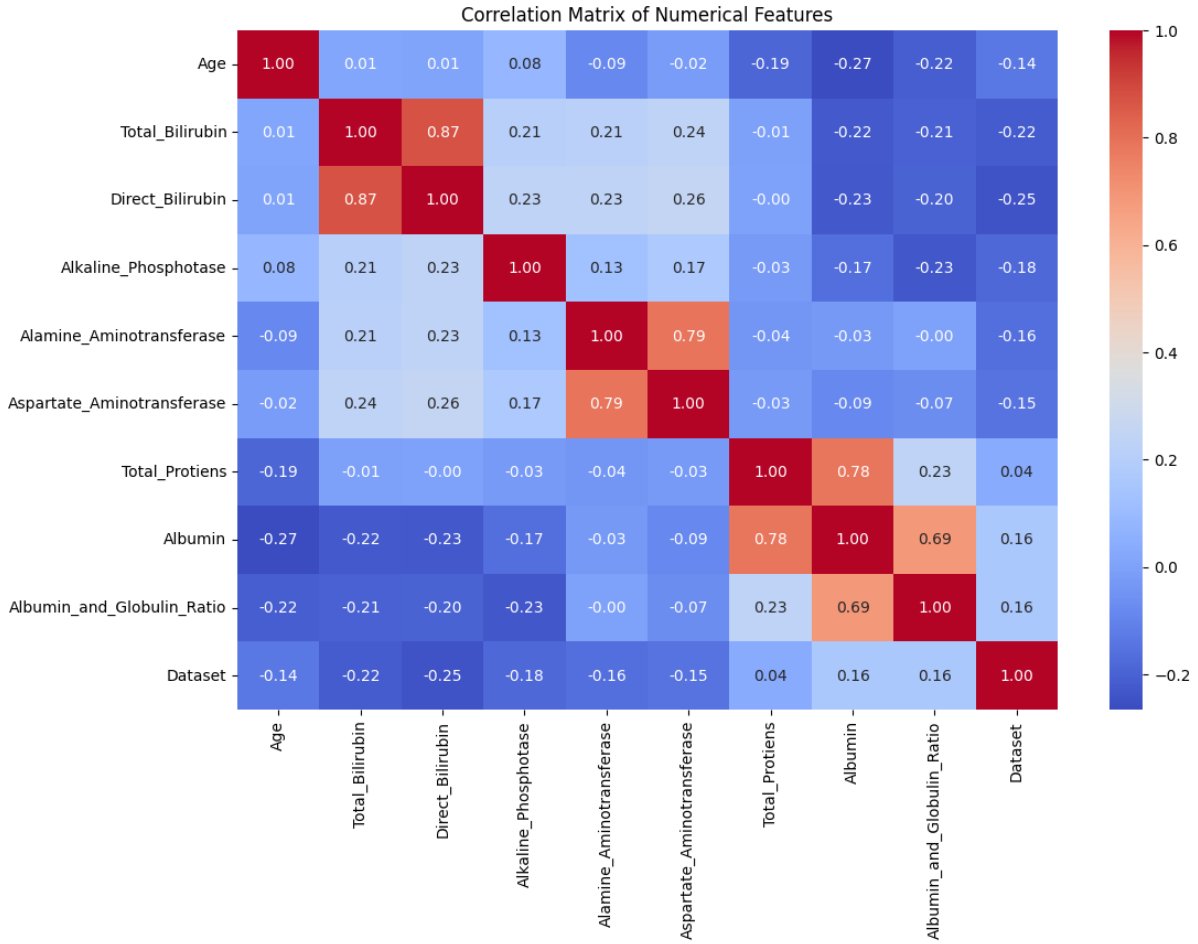

Figure 3.3.3 Correlation Matrix of Numeric Features

3.4 Model Selection

Choosing the right model is a vital part of creating an effective machine learning solution. This process involves selecting algorithms that best fit the characteristics of the dataset and the specific problem. Different models have their pros and cons; for instance, decision trees are easy to understand but may overfit the data, while support vector machines are effective for complex datasets but can require significant computational resources. It's important to test multiple models and assess their performance on validation data to find the most accurate one. The goal is to ensure that the chosen model performs well on new, unseen data, avoiding issues like overfitting or underfitting. Some models excel with high-dimensional data, while others are better for simpler or smaller datasets.

3.5 Training and Testing

The process involves two main phases - training the model on a dataset to learn the patterns and relationships, and then evaluating the trained model's performance on a separate test dataset that it hasn't seen before. This testing phase is critical to assess how well the model can generalize to new, unseen data, and identify any issues like overfitting or underfitting. Properly

splitting the available data into representative training and testing sets, and using techniques like cross-validation, are important best practices to ensure the model will perform well in real-world applications. Careful consideration of the training and testing process is essential for developing reliable and effective machine learning solutions.

3.6Model Evaluation

Evaluation metrics commonly used include accuracy, precision, recall, F1-score

a. Precision:

Precision is the proportion of true positive predictions to the total number of positive predictions (both true positives and false positives). It reflects how many of the predicted positive cases were genuinely accurate.

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

b. Recall:

Recall (sensitivity) is the proportion of true positive predictions to the total number of actual positive instances (true positives and false negatives). It demonstrates how many of the actual positive cases were accurately identified.

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

3.8 **c. F1-Score:**

The F1-score is the harmonic average of precision and recall, offering a single metric that balances both considerations.

$$F1 - score = 2 * \frac{(\text{precision} * \text{recall})}{(\text{precision}+\text{recall})}$$

d. Accuracy:

Accuracy is the proportion of correctly predicted instances (both true positives and true negatives) to the total number of instances. It reflects the overall correctness of the model.

$$accuracy = \frac{\text{number of correct predictions}}{\text{total number of predictions}}$$

CHAPTER-4
RESULTS AND DISCUSSION

4.1 Model Comparison

Decision Tree

The Decision Tree classifier, optimized using GridSearchCV, achieved an accuracy of 83.23% on the liver disease dataset, with the best parameters being Gini criterion, a max depth of 10, and minimum samples of 1 for both leaf nodes and splits. The model showed balanced performance between the two classes. For class 0 (No Liver Disease), it had a precision, recall, and F1-score of 81%, while for class 1 (Liver Disease), precision, recall, and F1-score were 85%. The overall classification performance is reflected in a macro and weighted average F1-score of 83%. The confusion matrix revealed 61 true negatives and 78 true positives, alongside 14 false positives and 14 false negatives, indicating the model's ability to effectively differentiate between liver disease and non-disease cases.

Random Forest

The classification report for the Random Forest model shows an overall accuracy of 92%, with precision, recall, and F1-score all being 0.92 for both classes. For class 0, the precision is0.95 and recall is 0.89, while for class 1, precision is 0.89 and recall is 0.95, indicating a good balance between the two classes. The macro and weighted averages for precision, recall, and F1-score are also 0.92. The model's best hyperparameters include 1350 estimators, a minimum of 3 samples to split a node, 1 sample per leaf, with a maximum depth of 40, and 'sqrt' as the maximum features to consider.

Artificial Neural Network

The classification report and confusion matrix for the Artificial Neural Network (ANN) model show an overall accuracy of 81.4%, with class 0 having a precision of 0.80, recall of 0.79, and F1-score of 0.79, while class 1 achieves slightly better results with a precision of 0.83, recall of 0.84, and F1-score of 0.83. The confusion matrix indicates that for class 0, 59 instances were correctly classified while 16 were misclassified, and for class 1, 77 were correctly classified while 15 were misclassified. The model has balanced performance across both classes, as reflected by the macro and weighted averages, both standing at 0.81 for precision, recall, and F1-score.

Ensemble Learning using Stack Classifier(Combines both svm and logistic regression)

This Algorithm performing is performing exceptionally well!

The stacking classifier, which combines the strengths of SVM and Logistic Regression, demonstrates exceptional performance with an impressive 95.00% accuracy. The model

achieves a balanced and consistent performance across both classes, with precision, recall, and F1-scores of 0.95 for each class, reflecting its ability to accurately predict both class 0 and class 1 instances without bias. The macro and weighted averages of 0.95 further highlight the model's robustness in handling varying class distributions. This well-rounded classifier proves its reliability, effectively balancing sensitivity and precision, making it a highly commendable choice binary classification tasks.

```
Stacking Classifier Accuracy for SVM and Logistic Regression: 95.00%
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.95      0.95       102
           1       0.95      0.95      0.95        98

    accuracy                           0.95       200
   macro avg       0.95      0.95      0.95       200
weighted avg       0.95      0.95      0.95       200
```

Model	Accuracy
Decision Tree	0.83
Random Forest	0.92
Artificial Neural Network	0.81
Ensemble Learning	0.95

Table 4.1: Models Comparison

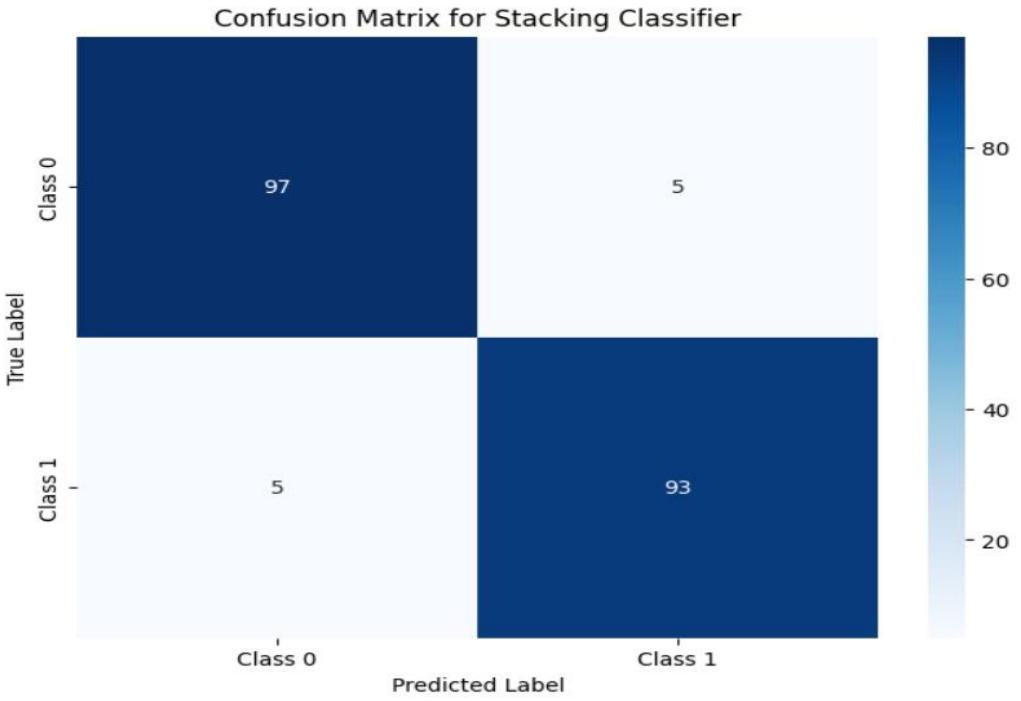

Fig 4.1 Confusion Matrix

CHAPTER-5

CONCLUSION

5.1 CONCLUSION

In summary, the comparative analysis of different machine learning models applied to the liver disease dataset yields significant insights into their effectiveness. The **Decision Tree** classifier achieved a commendable accuracy of **83.23%**, demonstrating balanced precision and recall across both classes. Despite its straightforward nature, the Decision Tree showed solid performance, with F1-scores of **81%** for class 0 (No Liver Disease) and **85%** for class 1 (Liver Disease), making it a practical choice for quick and interpretable assessments. The confusion matrix indicated a reasonable classification accuracy, although a few misclassifications were noted. Overall, while the Decision Tree performed adequately, more advanced models in the analysis surpassed its results.

In contrast, the **Random Forest** classifier significantly exceeded the Decision Tree's performance, achieving an accuracy of **92%**. It maintained a well-balanced performance across both classes, with precision, recall, and F1-scores all reaching **0.92**. This highlights the Random Forest's capability to handle complex datasets effectively by minimizing variance and preventing overfitting, particularly through its use of numerous estimators and optimized hyperparameters. The model adeptly captured the intricacies of the data, as indicated by the balanced error distribution in the confusion matrix. Nevertheless, while the Random Forest improved upon the Decision Tree, it was not the highest-performing model in this analysis.

The standout performer was the **Stacking Classifier**, which combines **SVM** and **Logistic Regression**, achieving an impressive **95% accuracy**. This ensemble method leverages the strengths of its constituent algorithms, resulting in precision, recall, and F1-scores of **0.95** for both classes, surpassing all other models. The Stacking Classifier exhibited exceptional balance in its predictions, effectively minimizing both false positives and false negatives. Its outstanding performance across various metrics, along with its capacity to generalize well to unseen data, positions it as the most dependable option for binary classification tasks like predicting liver disease. The results suggest that integrating multiple algorithms can lead to substantial enhancements in prediction accuracy, making the Stacking Classifier the most noteworthy model in this evaluation.

REFERENCES

[1] S. S. Nigatu, P. C. R. Alla, R. N. Ravikumar, K. Mishra, G. Komala, and G. R. Chami, "A Comparitive Study on Liver Disease Prediction using Supervised Learning

Algorithms with Hyperparameter Tuning," *2023 Int. Conf. Adv. Comput. Comput. Technol. InCACCT 2023*, pp. 353–357, 2023, doi: 10.1109/InCACCT57535.2023.10141830.

[2] A. E. Topcu, E. Elbasi, and Y. I. Alzoubi, "Machine Learning-Based Analysis and Prediction of Liver Cirrhosis," *2024 47th Int. Conf. Telecommun. Signal Process. TSP 2024*, pp. 191–194, 2024, doi: 10.1109/TSP63128.2024.10605929.

[3] C. Anuradha, D. Swapna, B. Thati, V. N. Sree, and S. P. Praveen, "Diagnosing for Liver Disease Prediction in Patients using Combined Machine Learning Models," *Proc. - 4th Int. Conf. Smart Syst. Inven. Technol. ICSSIT 2022*, pp. 889–896, 2022, doi: 10.1109/ICSSIT53264.2022.9716312.

[4] A. J. M. Rani, S. Nishanthini, D. C. J. Josephine, H. Venugopal, S. G. Nissi, and V. Jacintha, "Liver Disease Prediction using Semi Supervised based Machine Learning Algorithm," *3rd Int. Conf. Smart Electron. Commun. ICOSEC 2022 - Proc.*, no. Icosec, pp. 1389–1392, 2022, doi: 10.1109/ICOSEC54921.2022.9952144.

[5] A. Dhyani *et al.*, "Comparative Analysis of Supervised Machine Learning Algorithms for Liver Disease Prediction with SMOTE Enhancement," *2023 3rd Asian Conf. Innov. Technol. ASIANCON 2023*, pp. 1–6, 2023, doi: 10.1109/ASIANCON58793.2023.10270381.

[6] M. Minnoor and V. Baths, "Liver Disease Diagnosis Using Machine Learning," *Proc. - 2022 IEEE World Conf. Appl. Intell. Comput. AIC 2022*, pp. 41–47, 2022, doi: 10.1109/AIC55036.2022.9848916.

[7] H. S. Yadav and R. K. Singhal, "Classification and Prediction of Liver Disease Diagnosis Using Machine Learning Algorithms," *2023 2nd Int. Conf. Innov. Technol. INOCON 2023*, pp. 1–6, 2023, doi: 10.1109/INOCON57975.2023.10101221.

[8] S. Muhuri, A. Sarkar, S. Chakraborty, and S. Chakraborty, "A Statistical Method for Prediction of Liver Disease based on the Brownian Motion Model," *Proc. 2019 IEEE Reg. 10 Symp. TENSYMP 2019*, vol. 7, pp. 157–161, 2019, doi: 10.1109/TENSYMP46218.2019.8971234.

[9] S. Ambesange, A. Vijayalaxmi, R. Uppin, S. Patil, and V. Patil, "Optimizing Liver disease prediction with Random Forest by various Data balancing Techniques," *Proc. - 2020 IEEE Int. Conf. Cloud Comput. Emerg. Mark. CCEM 2020*, pp. 98–102, 2020, doi: 10.1109/CCEM50674.2020.00030.

[10] S. Ambesange, R. Nadagoudar, R. Uppin, V. Patil, S. Patil, and S. Patil, "Liver Diseases Prediction using KNN with Hyper Parameter Tuning Techniques," *Proc. B-HTC 2020 - 1st IEEE Bangalore Humanit. Technol. Conf.*, pp. 1–6, 2020, doi: 10.1109/B-HTC50970.2020.9297949.

[11] B. H. Al Telaq and N. Hewahi, "Prediction of Liver Disease using Machine Learning Models with PCA," *2021 Int. Conf. Data Anal. Bus. Ind. ICDABI 2021*, pp. 250–254, 2021, doi: 10.1109/ICDABI53623.2021.9655897.

[12] S. R. Tanuku, A. A. Kumar, S. R. Somaraju, R. Dattuluri, M. V. K. Reddy, and S. Jain,

"Liver Disease Prediction Using Ensemble Technique," *8th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2022*, vol. 1, pp. 1522–1525, 2022, doi: 10.1109/ICACCS54159.2022.9784999.

[13] R. Zhao, X. Wen, H. Pang, and Z. Ma, "Liver disease prediction using W-LR-XGB Algorithm," *Proc. - 2021 Int. Conf. Comput. Blockchain Financ. Dev. CBFD 2021*, pp. 245–248, 2021, doi: 10.1109/CBFD52659.2021.00055.

[14] S. Kumar and P. Rani, "A Comparative Survey on Machine Learning Techniques for Prediction of Liver Disease," *Proc. Int. Conf. Contemp. Comput. Informatics, IC3I 2023*, vol. 6, pp. 1796–1801, 2023, doi: 10.1109/IC3I59117.2023.10397980.

[15] M. A. Kuzhippallil, C. Joseph, and A. Kannan, "Comparative Analysis of Machine Learning Techniques for Indian Liver Disease Patients," *2020 6th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2020*, pp. 778–782, 2020, doi: 10.1109/ICACCS48705.2020.9074368.

[16] A. Rout, S. Mishra, V. Sharma, O. D. M. Chiadika, T. T. Tonukari, and C. Iwendi, "Non-Alcoholic Fatty Liver Disease Prediction with Feature Optimized XGBoost Model," *4th Int. Conf. Innov. Pract. Technol. Manag. 2024, ICIPTM 2024*, no. Iciptm, pp. 1–5, 2024, doi: 10.1109/ICIPTM59628.2024.10563859.

[17] Y. Xiao, "Machine Learning Based Liver Cancer Disease Prediction System Using Improved Extreme Gradient Boosting Algorithm," *2024 Second Int. Conf. Data Sci. Inf. Syst.*, pp. 1–5, 2024, doi: 10.1109/icdsis61070.2024.10594135.

[18] K. Prakash and S. Saradha, "Intelligent MRI Liver Images based Cirrhosis Disease Identification using Modified Learning Principle," *Int. Conf. Edge Comput. Appl. ICECAA 2022 - Proc.*, no. Icecaa, pp. 1391–1398, 2022, doi: 10.1109/ICECAA55415.2022.9936406.

[19] R. T. Umbare, O. Ashtekar, A. Nikhal, B. Pagar, and O. Zare, "Prediction and Detection of Liver Diseases using Machine Learning," *3rd IEEE Int. Conf. Technol. Eng. Manag. Soc. Impact using Mark. Entrep. Talent. TEMSMET 2023*, pp. 1–6, 2023, doi: 10.1109/TEMSMET56707.2023.10150135.

[20] S. M. Ganie, P. K. Dutta Pramanik, and Z. Zhao, "Improved liver disease prediction from clinical data through an evaluation of ensemble learning approaches," *BMC Med. Inform. Decis. Mak.*, vol. 24, no. 1, pp. 1–24, 2024, doi: 10.1186/s12911-024-02550-y.