A Minor Project Report
On
MACHINE LEARNING MODEL FOR EARLIER PREDICTION OF LIVER DISEASE

A Dissertation Submitted
In Partial Fulfillment of the Requirements for the Award of the Degree of

Bachelor of Technology
In
COMPUTER SCIENCE AND ENGINEERING
(ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)

By

VIJAYAGIRI SWETHA 22211A66C2

Under the Esteemed Guidance of
Dr. Hutashan Vishal Bhagat and Mrs. B. Lavanya

VISHNU
UNIVERSAL LEARNING

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING (ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)

B V RAJU INSTITUTE OF TECHNOLOGY
(UGC Autonomous)
Vishnupur, Narsapur, Medak (District) – 502313, TS
(Affiliated to JNTUH and Approved by AICTE)
2023-2024

DECLARATION

We here by declare that the Minor Project Report entitled **"Machine Learning Model For Earlier Prediction Of Liver Disease "** submitted to the Department of Computer Science and Engineering (Artificial Intillegence and Machine Learning), B V Raju Institute of Technology, in partial fulfillment of the requirements for the Award of Degree of Bachelor of Technology in Computer Science and Engineering (Artificial Intelligence and Machine Learning) is Our cord of the original work done by us and has not been submitted to any institute or published elsewhere.

Place: Narsapur

Date:

Signature

Vijayagiri Swetha **22211A66C2**

B V RAJU INSTITUTE OF TECHNOLOGY

(UGC Autonomous, Accredited by NBA & NAAC)

Vishnupur, Narsapur, Medak, Telangana, India – 502313

CERTIFICATE

This is to certify that the minor project entitled "**MACHINE LEARNING MODEL FOR EARLIER PREDICTION OF LIVER DISEASE**" being submitted by

VIJAYAGIRI SWETHA **22211A66C2**

In Partial fulfilment of requirements for the award of degree of Bachelor of Technology in **Computer Science and Engineering (Artificial Intelligence and Machine Learning)** to **B V Raju Institute of Technology** is a record of bonafide work carried out during the period of February 2024 to June 2024 by them under the supervision of

Dr. Hutashan Vishal Bhagat **Supervisor**

Mrs. B Lavanya **Co-Supervisor**

This is to certify that the above statement made by the students are correct to the best of our knowledge.

Mrs. B. Lavanya **Dr. Hutashan Vishal Bhagat**

Co-Supervisor **Supervisor**

The Minor Project Viva-Voce for this team has been held on _____.

External Examiner **Dr. G. Uday Kiran**

Program Coordinator

ACKNOWLEDGEMENT

This acknowledgement transcends the reality of formality where We would express deep gratitude and respect to all those people who helped supported, guided and inspired us throughout the completion of this Minor Project.

We wish to convey our sincere thanks to our Supervisor, D**r. Hutashan Vishal Bhagat, Assistant Professor** and **Mrs. B Lavanya, Assistant Professor Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning), B V Raju Institute of Technology** for their continuous support.

We wish to convey our sincere thanks to **Dr. G Uday Kiran, Associate Professor and Program Coordinator, Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning), B V Raju Institute of Technology** for his continuous support.

We also grateful to all Teaching and Non-Teaching Faculty of Computer Science and Engineering (Artificial Intelligence and Machine Learning) Department, people who provided us the useful support needed for this Minor Project. Lastly, We thank all our Family, Friends and Well - Wishers.

VIJAYAGIRI SWETHA 22211A66C2

ABSTRACT

This project aims to develop a machine learning model for predicting liver disease through a thorough analysis of patient data. The workflow employs various libraries, including NumPy and Pandas for managing data, Matplotlib and Seaborn for creating visualizations, and Scikit-learn for implementing machine learning algorithms. The data preprocessing involves handling missing values with KNN Imputer, normalizing data using StandardScaler, and addressing class imbalance with SMOTE.

Several classification algorithms are explored, such as Random Forest, Logistic Regression, K-Nearest Neighbors, Support Vector Classifier, and Gaussian Naive Bayes. A Voting Classifier is also utilized to aggregate predictions from these models. To optimize performance, hyperparameter tuning is carried out using Optuna. The dataset, which comprises liver function test results and demographic information, is carefully processed to enhance the model's accuracy.

Model evaluation is conducted using metrics like Accuracy, Precision, Recall, and F1 Score. Additionally, SHAP values are employed to interpret the model's predictions and analyze the impact of different features. This project provides a holistic approach to liver disease prediction by integrating effective data preprocessing, model optimization, and comprehensive evaluation to achieve precise and interpretable outcomes.

Key Words: Voting Classifier, Optuna, SHAP.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER-1
INTRODUCTION

1.1 Introduction of Liver Diseases

Liver issues can be difficult to identify early because patients may continue to operate normally despite having partial liver damage. Early detection is crucial, as it greatly enhances the likelihood of survival for individuals with liver disease. The liver plays a critical role in various bodily functions, including energy storage, metabolism, and detoxification. It assists in digesting food, transforming it into energy, and storing energy for future use. Moreover, the liver helps eliminate harmful substances from the blood. The term "liver disease" encompasses a range of conditions that impact the liver's health and function.

Functions of the Liver

The liver carries out several essential functions:

1. It generates components of the immune system to fight infections.
2. It synthesizes proteins that help with blood clotting.
3. It breaks down old or damaged red blood cells.
4. It stores surplus blood sugar in the form of glycogen.

A variety of disorders can affect the liver and disrupt its functions. While some conditions can be effectively treated, others may be more resistant to treatment.

Common Liver Conditions

Autoimmune Hepatitis: This condition arises when the immune system mistakenly attacks and destroys healthy liver tissue, which can lead to cirrhosis and other types of liver damage.

1. **Cirrhosis**: Cirrhosis is characterized by the replacement of healthy liver tissue with scar tissue. This often results from chronic hepatitis, prolonged excessive alcohol use, or rare genetic disorders such as Wilson's disease.
2. **Hemochromatosis**: This disorder involves an abnormal accumulation of iron in the body, which can cause damage to the liver.
3. **Hepatitis A**: Hepatitis A is a viral infection that causes inflammation of the liver. It is commonly found in areas with inadequate sanitation and limited access to clean water.
4. **Hepatitis B**: Hepatitis B can be acute or chronic and is transmitted through contact with infectious bodily fluids, such as via shared needles or accidental needle sticks. It can lead to severe complications like liver failure and cancer, though vaccination is available for prevention.
5. **Hepatitis C**: Hepatitis C is a viral infection that can be either acute or chronic, typically spread through exposure to infected blood, such as from contaminated needles. It may lead to liver failure and liver cancer.
6. **Non-Alcoholic Fatty Liver Disease (NAFLD) and Non-Alcoholic Steatohepatitis (NASH)**: These conditions involve the accumulation of excess fat in the liver, leading to inflammation and damage. NAFLD may progress to fibrosis, while NASH can result in severe liver damage and is often associated with type 2 diabetes.

Figure 1.1 Stages Of Liver

Symptoms of Liver Conditions

Liver disorders can initially present with flu-like symptoms and may progress to more severe manifestations, such as jaundice and dark urine. Common symptoms of liver disease include:

- Fatigue
- Reduced appetite
- Nausea and vomiting
- Joint pain
- Abdominal pain or discomfort
- Nosebleeds
- Unusual blood vessels on the skin (spider angiomas)

More severe symptoms may include:

- Yellowing of the skin and eyes (jaundice)
- Abdominal swelling (ascites)
- Swollen legs (edema)
- Enlargement of breasts in men (gynecomastia)
- Enlarged liver (hepatomegaly)

1.2 Problem Statement:

Liver diseases, including cirrhosis, hepatitis, and fatty liver disease, represent a major global health issue. These conditions often develop without noticeable symptoms, which can lead to severe health consequences or even death if not identified and treated early. The process of diagnosing liver disease is often complicated by intricate diagnostic procedures and insufficient medical

resources, leading to lengthy wait times and reduced effectiveness in early detection. Consequently, delays in diagnosis can result in late-stage treatment and adverse outcomes for patients.

This project aims to tackle these issues by creating a machine learning-based predictive model for diagnosing liver disease using data derived from blood tests. The objective is to develop a highly accurate and reliable system that combines sophisticated data preprocessing, various classification algorithms, and ensemble learning methods to facilitate early diagnosis. By employing algorithms such as Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Random Forest, and Naive Bayes, and fine-tuning model performance through Optuna's hyperparameter optimization, this project seeks to improve diagnostic precision and efficiency.

1.3 Project Scope

The liver is the largest internal organ in the human body and plays a vital role in digesting food, detoxifying harmful substances, and supporting metabolism along with other critical functions. This study is driven by the rising prevalence of liver disease globally, which poses a significant health challenge due to its high mortality rate. Early detection of liver diseases can greatly reduce complications and enhance patient outcomes. This research seeks to utilize advanced technologies to overcome delays in diagnosing and treating liver cirrhosis, thereby alleviating the strain on healthcare systems.

1.4 Project Purpose

The main objective of this project is to improve the early detection of liver disease, which is often hindered by long wait times due to limited medical resources and

intricate diagnostic procedures. Liver disease is a major global health issue, with liver cirrhosis being a prominent concern due to its gradual progression. Early detection is crucial for enhancing survival rates. This project aims to minimize diagnostic delays, thereby facilitating timely intervention and making the treatment of liver cirrhosis more cost-effective. By enabling healthcare professionals to act promptly and efficiently, the project seeks to enhance patient outcomes.

1.5 Expected Outcomes

- A highly accurate and reliable predictive model for diagnosing liver diseases based on blood test parameters.
- An in-depth analysis of feature importance, illustrating how various blood test results affect liver disease diagnosis.
- An effective and automated method for hyperparameter optimization, ensuring optimal model performance.
- Clear and interpretable visualizations of model predictions to assist medical professionals in understanding and trusting the model's decisions.

Significance

This project aims to deliver a non-invasive, precise, and efficient diagnostic tool for liver diseases, which could lead to earlier diagnosis, improved patient outcomes, and reduced healthcare expenses. By employing advanced machine learning techniques and addressing key challenges such as class imbalance and

hyperparameter optimization, the project seeks to offer valuable insights and tools for advancing medical diagnostics.

1.6 Algorithm

1.6.1. Logistic Regression

Overview: Logistic Regression is a straightforward linear model used primarily for binary classification. It estimates the probability of a binary outcome using a logistic function that transforms any real-valued input into a range between 0 and 1.

How it Works:

- Logistic Regression computes a weighted sum of the input features.
- It then applies a logistic (sigmoid) function to convert this sum into a probability between 0 and 1.
- A threshold (usually 0.5) is used to classify the input into one of the two possible classes.

Strengths:

- Easy to implement and interpret.
- Provides probability estimates that are useful for understanding the model's confidence.
- Can be regularized with L1 or L2 techniques to prevent overfitting.

Role in the Project: Logistic Regression offers a baseline performance and helps in interpreting the effect of each blood test parameter on the likelihood of liver

disease. Its simplicity and interpretability make it a useful starting point for analyzing the data.

1.6.2 K-Nearest Neighbors (KNN)

Overview: K-Nearest Neighbors (KNN) is an instance-based, non-parametric algorithm used for both classification and regression. It classifies a data point based on the majority class among its k-nearest neighbors.

How it Works:

- The algorithm retains all training data points.
- For a new data point, it calculates the distance to all existing training points (usually with Euclidean distance).
- It identifies the k-nearest neighbors and assigns the class based on the majority vote.

Strengths:

- Intuitive and simple to understand.
- Effective for smaller datasets or those with clear decision boundaries.
- Does not require a training phase, making it fast for training.

Role in the Project: KNN can capture local patterns in the data, which is beneficial for identifying specific trends in liver disease based on blood test results. It provides a clear understanding of how similar patients are classified based on proximity in the feature space.

1.6.3. Support Vector Machine (SVM)

Overview: Support Vector Machine (SVM) is a robust classification technique that finds the optimal hyperplane to separate different classes in the feature space, maximizing the margin between them.

How it Works:

- SVM identifies the hyperplane that maximizes the margin between the nearest points of different classes (support vectors).
- It can utilize kernel functions (such as linear, polynomial, or RBF) to handle non-linear boundaries.

Strengths:

- Effective in high-dimensional spaces.
- Resilient to overfitting, particularly in high-dimensional contexts.
- Capable of modeling non-linear relationships through kernel functions.

Role in the Project: SVM is instrumental in detecting complex, non-linear relationships between blood test parameters and liver disease. Its ability to maximize class separation helps achieve high accuracy and reliability in predictions.

1.6.4. Random Forest

Overview: Random Forest is an ensemble learning method that constructs multiple decision trees during training and combines their outputs to enhance accuracy and minimize overfitting.

How it Works:

- It generates multiple decision trees using different subsets of training data and features (bootstrap sampling and feature sampling).

- Each tree provides a prediction, and the final output is the mode of these predictions for classification tasks.

Strengths:

- Handles large and high-dimensional datasets effectively.
- Reduces overfitting through averaging across multiple trees.
- Offers feature importance metrics to highlight the most predictive features.

Role in the Project: Random Forest is valuable for capturing intricate interactions between blood test parameters. It helps identify the most significant features for liver disease diagnosis and provides robust performance even with noisy or irrelevant features.

1.6.5. Naive Bayes (GaussianNB)

Overview: Naive Bayes is a probabilistic classifier based on Bayes' theorem, which assumes feature independence. Gaussian Naive Bayes (GaussianNB) specifically assumes that features follow a normal distribution.

How it Works:

- It computes the posterior probability for each class given the input features.
- Bayes' theorem is used to combine prior probabilities with feature likelihoods.
- The class with the highest posterior probability is selected.

Strengths:

- Simple and computationally efficient.

- Performs well with high-dimensional data.
- Provides probabilistic predictions.

Role in the Project: Naive Bayes acts as a simple yet effective baseline model. Its probabilistic approach aids in understanding the likelihood of liver disease given the blood test parameters. Despite its simplicity, it can be effective, especially when the independence assumption approximately holds.

1.6.6 Ensemble Learning

Overview: Ensemble learning is a method in machine learning where multiple models, often referred to as "weak learners," are combined to create a single, more powerful model. The core concept is that while individual models might have limitations, their combined output can improve performance, reduce errors, and enhance overall robustness.

Voting Classifiers

Overview: A Voting Classifier is a specific type of ensemble method used for classification tasks. It integrates predictions from various classifiers to boost overall prediction accuracy. The aim is to harness the diverse strengths of different classifiers to build a more accurate and reliable model.

Types of Voting:

- **Hard Voting:**
    - Each classifier provides a vote for a particular class label.
    - The final prediction is the class that receives the majority of votes.
- **Soft Voting:**
    - Each classifier provides a probability estimate for each class.

- The final prediction is the class with the highest average probability across all classifiers.

Benefits of Voting Classifiers:

- **Enhanced Accuracy:** By aggregating predictions from multiple models, a Voting Classifier often achieves better accuracy than any single model alone.
- **Increased Robustness:** Voting classifiers are less prone to overfitting as they average out the errors across several models.
- **Flexibility:** Voting classifiers can incorporate various types of models (e.g., logistic regression, KNN, random forest), utilizing their complementary strengths to improve performance.

Role in the Project: In the context of liver disease diagnosis, a Voting Classifier can integrate predictions from different models, such as Logistic Regression, KNN, SVM, Random Forest, and Naive Bayes. This approach helps to:

- **Utilize Diverse Strengths:** Combining different models allows the ensemble to leverage each model's unique strengths, compensating for individual weaknesses.
- **Minimize Variance:** Averaging the predictions from multiple models helps to reduce variance, leading to more consistent and reliable results.
- **Boost Performance:** The ensemble approach is likely to offer improved performance over any single model, as evidenced by the high accuracy achieved on both validation and test datasets.

In summary, ensemble learning, and particularly Voting Classifiers, are essential for enhancing the precision and stability of the predictive model used for diagnosing liver diseases from blood test parameters.

CHAPTER-2

LITERATURE REVIEW

Souvik Sarkar et al. [1] used the UC Davis OMOP Database and UCSF Information Commons OMOP Database. They processed data in the KNIME Analytics Platform, encoding categorical data for machine learning algorithms. They tested Random Forest (RF), Gradient Boosting (GB), Naïve Bayes (NB), Stochastic Gradient Descent, K-Nearest Neighbor (KNN), and Probabilistic Neural Network (PNN). The Gradient Boosted model achieved the highest performance with 92.12% accuracy and a 0.97 AUC, followed by the Random Forest model with 90.29% accuracy and a 0.97 AUC.

Jose Luis Calleja et al. [2] used gene expression data to identify a 10-gene signature for liver cancer stem cells. They employed XGBoost for feature ranking and SHAP values for feature impact analysis, achieving 97% accuracy. This signature shows promise as a biomarker for liver cancer detection and characterization.

Qichen Chen et al. [3] used training and external validation datasets with patient characteristics. They employed data imputation integrated into the Random Forest (RF) model and conducted all analyses using R. The algorithms used were Random Forest (RF), Random Survival Forest (RSF), cross-validation, performance metrics, and decision curve analysis (DCA). The RF and RSF models successfully predicted clinical outcomes for CRLM patients, significantly improving over traditional risk scores and demonstrating practical clinical utility.

Bingyu Wang et al. [4] used the GSE89632 dataset and various R packages to construct diagnostic models. The authors employed ROC curves to assess predictive efficacy and used Random Forest (RF), Support Vector Machine Recursive Feature Elimination (SVM-RFE), and LASSO Regression algorithms. The study identified BCL2L11, NAGS, HDHD3, and RMND1 as potential diagnostic biomarkers for NAFLD, validated their predictive efficacy, and developed a diagnostic model with significant clinical benefit.

The study by Eloy Ruiz et al. [5] involved 263 patients with non-cirrhotic liver hepatocellular carcinoma (NCL-HCC) who underwent liver resection. Data management utilized Excel and PostgreSQL, with statistical analysis performed using R (v4.2.1) and Stata (v14.0). They applied the Kaplan-Meier method and Random Forest algorithm from the R package to distinguish early and late recurrences, achieving significant insights through a minimum P-value approach combined with random survival forest modelling.

Ahmed M. Elshewey et al. [6] utilized a dataset with 1385 records and 29 attributes from the UCI repository. They applied Min-Max normalization and forward selection for feature selection. Their main algorithm was an optimized gradient boosting (GB) classifier, fine-tuned with the OPTUNA framework. Comparatively analyzed against decision tree (DT), support vector machine (SVM), dummy classifier (DC), ridge classifier (RC), and bagging classifier (BC), the optimized GB model achieved superior performance with a 95.3% accuracy, attributed to the OPTUNA framework's optimized hyperparameters.

Mandakini Priyadarshani Behera et al. [7] used a dataset from the UCI Machine Learning Repository. They employed machine learning techniques, focusing on

classification algorithms, particularly a hybrid model combining Support Vector Machine (SVM) with a modified Particle Swarm Optimization (PSO). The study tested SVM, modified PSO, PSOSVM, CPSOSVM, and CCPSOSVM algorithms. Results showed that the hybrid model, especially CCPSOSVM, achieved the highest classification accuracy and lowest error rate for heart and liver disease prediction, evaluated by accuracy, error rate, precision, recall, and F1 score.

Harish Gadhe et al. [8] used CT scan images from OASIS and TCIA datasets to predict liver diseases. They applied noise removal, contrast enhancement, resizing, and data augmentation. A Convolutional Neural Network (CNN) with ReLU activation, pooling layers, SoftMax activation, dropout regularization, and batch normalization was employed. The dataset was split into 80% training and 20% validation. The CNN, optimized using binary cross-entropy and stochastic gradient descent, achieved high accuracy. Performance was validated with accuracy, precision, recall, F1 score, and AUC metrics.

Ruhul Amin et al. [9] used the Indian Liver Patient Dataset (ILPD) from the UCI repository, which includes 583 records. They applied data pre-processing, including imputation, outlier handling, and class imbalance correction. Feature extraction used PCA, FA, and LDA. They tested Logistic Regression, Random Forest, KNN, SVM, MLP, and an Ensemble Voting Classifier. Results: 88.10% accuracy, 85.33% precision, 92.30% recall, 88.68% F1 score, and 88.20% AUC. Their method significantly improves liver disease classification, aiding in diagnosis and potentially improving patient outcomes.

Zainab Sattar Jabbar et al. [10] used 700 ultrasound shear wave elastography images, categorized into five classes, split into 90% training and 10% testing

sets. They applied image processing techniques and used a Convolutional Neural Network (CNN) for feature extraction, followed by classification using SoftMax and Support Vector Machine (SVM). The CNN-SVM classifier achieved 98.59% accuracy, outperforming the CNN-SoftMax classifier's 97.18%, with higher precision, recall, and F1-score.

Valeriu Harabor et al. [11] conducted a prospective cohort screening study in Romania from January to November 2022. Using feature selection, they trained models to predict Hepatitis C (HCV) and Hepatitis B (HBV) status. Algorithms used included SVM, Random Forest (RF), Naive Bayes (NB), and K-Nearest Neighbors (KNN). For HCV prediction, KNN achieved the highest accuracy of 98.1%, SVM and RF both reached 97.6%, and NB achieved 95.7%. HBV prediction accuracy ranged from 78.2% to 97.6%, showing better performance for HCV prediction overall.

Zhengyun Zhao et al. [12] utilized data from 998 Hepatocellular Carcinoma (HCC) patients spanning 2010 to 2018. They employed Random Survival Forests (RSF), B-splines, and Hierarchical Clustering for liver cancer staging. Their novel RSF-based model surpassed the BCLC system, demonstrating enhanced accuracy in survival prediction and increased clinical relevance.

Umesh Kumar Lilhore et al. [13] used the online HCV dataset from UCI to measure the performance of their proposed Hybrid Predictive Model (HPM) for precise hepatitis-C classification. They handled data imbalance using the Synthetic Minority Over-Sampling Technique (SMOTE). The authors performed feature selection using a Ranker method and applied improved random forest (IRF) with support vector machine (SVM) to select higher-ranked features for

building the prediction model. They conducted two experiments: one based on dataset splitting methods (K-fold cross-validation and training-testing split) and another on feature selection with and without SMOTE. The proposed method achieved an accuracy of 96.82% with SMOTE-based feature selection. The study demonstrated the importance of feature selection and the effectiveness of the HPM model in achieving higher accuracy compared to existing methods

Lukas Otero Sanchez et al. [14] used a prospective cohort of patients with type 1 or type 2 diabetes without advanced fibrosis. They employed a novel machine learning method for diabetic cluster identification, revealing patients with severe insulin resistance at high risk of liver-related outcomes and fibrosis progression. They identified alcohol consumption at diagnosis as a significant risk factor for liver-related events.

Tahereh Jafari et al. [15] developed a Support Vector Machine (SVM) model to predict neonatal jaundice within the first 24-72 hours post-delivery. Using a Gaussian kernel with a sigma of 1.2360605, the model was tested on 354 cases, correctly predicting 321. Jaundice presence was the output variable, with 25 predictive factors. The SVM model outperformed other algorithms in classification precision, demonstrating its effectiveness for early detection and treatment of neonatal jaundice.

Stanislav Listopad et al. [16] used machine learning methods on gene expression data to classify alcohol-associated and non-alcohol-associated liver diseases. Gene expression profiling of peripheral blood mononuclear cells (PBMCs) helped characterize HBV, HCV, and primary biliary cholangitis, complementing liver tissue pathology. The study involved 137 PBMC and 67 liver tissue samples from

the Southern California Alcoholic Hepatitis Consortium. Diseases included alcohol-associated hepatitis, alcohol-associated cirrhosis, NAFLD, chronic HCV, and healthy controls. Using the GSE142530 dataset, differential expression and information gain methods selected features, with classifiers evaluated via k-fold nested cross-validation, achieving 75% overall accuracy.

Grace Lai-Hung Wong et al. [17] developed machine learning models to predict hepatocellular carcinoma (HCC) in patients with chronic viral hepatitis (CVH) using Hospital Authority Data Collaboration Lab (HADCL) data. They utilized ICD-9CM codes and aimed to create prediction models based on clinical and laboratory parameters. Using SPSS, SAS, and R software, they compared logistic regression, ridge regression, AdaBoost, decision tree, and random forest methods. Random forest achieved 90% accuracy, decision tree 80%, and ridge regression 85%. Ridge regression consistently showed high accuracy (90%) in the validation cohort. The study identified risk factors for HCC in patients with advanced liver fibrosis.

P.R.Kshirsagar et al. [18] used classification algorithms like SVM, K-means clustering, KNN, Random Forest, and Logistic Regression. With liver disease causing 3.5% of global deaths and rising mortality rates, early detection through AI can significantly improve patient outcomes. The research involves pre-processing, feature extraction, and classification, proposing a hybrid classification system for more accurate liver disease prediction.

Tzuu-Hseng et al. [19] uses the HCV dataset from the UCI Machine Learning Repository, containing 582 instances with 14 attributes. The technologies and algorithms used include Random Forest (RF), Logistic Regression (LR), and the Artificial Bee Colony (ABC) algorithm, along with SMOTE for handling imbalanced

data. The model achieved a prediction accuracy of 71.53%, surpassing other methods in weighted-average recall and F1 scores

Oladosu Oyebisi Oladimeji et al. [20] used the Hepatitis C Virus (HCV) dataset from the UCI Machine Learning Repository for their study. The dataset consists of Hepatitis C test records of 615 patients, including 238 women and 377 men, with an age range of 19 to 77 years. It contains 13 features related to various Hepatitis tests. The algorithms tested included Decision Tree (DT), K-Nearest Neighbors (KNN), Random Forest (RF), Naïve Bayes (NB), and Logistic Regression (LR). The best performing algorithm was Random Forest (RF), achieving an accuracy of 99% after applying SMOTE.

Mr. G Ragu et al. [21] studies on the dataset from the UCI Machine Learning Repository, consisting of 19 parameters related to Hepatitis B. Technologies and algorithms employed include data preprocessing, exploratory data analysis (EDA), feature selection (Select K-best and Recursive Feature Elimination), and a decision tree algorithm. The accuracy of the decision tree model is 74.4%, as validated by a confusion matrix. The results indicate that the decision tree is effective in predicting the living status of Hepatitis B patients.

Mohammed Alghobiri et al. [22] used the Liver disease dataset sourced from Kaggle for their study. The dataset contains over 550 records with 10 features.The algorithms tested included Naïve Bayes, Decision Tree, K-Nearest Neighbor (KNN), and Logistic Regression. The best performing algorithm was Decision Tree, achieving an accuracy of 72.5%.

Ashfaq Ali Kashif et al. [23] used a dataset collected from a hospital in Lahore, Pakistan, for their study. The dataset included various attributes related to the medical information of the patients, with the target attribute being the treatment

response categorized as "Respondent" or "Not Respondent" to LOLA therapy.The algorithms tested for prediction included K Nearest Neighbor, kStar, Naive Bayes, Random Forest, Radial Basis Function, PART, Decision Tree, OneR, Support Vector Machine, and Multi-Layer Perceptron. The best performing algorithm was Decision Tree (DT), achieving an accuracy of 85.9155%.

Boggarapu Sai Surya et al. [24] utilized the Hepatitis C Virus (HCV) dataset from the UCI Machine Learning Repository. The dataset contained 615 instances and 14 attributes, with 582 instances used after removing missing values. The technologies employed included feature selection, SMOTE for data balancing, and machine learning algorithms such as Random Forest (RF), Logistic Regression (LR), and the Artificial Bee Colony (ABC) algorithm for optimizing confidence thresholds. The proposed model, a combination of RF and LR, achieved an accuracy of 71.53%, surpassing other methods by up to 10.33% in various performance metrics .

Naresh Kumar Trivedi et al. [25] discussed the application of machine learning and deep learning for diagnosing COVID-19 through X-ray image analysis. The study utilized the Kaggle COVID-19 X-ray dataset, consisting of 600 normal images (250 for training, 350 for testing) and 5,000 COVID-19 images (2,000 for training, 3,000 for testing). Two methods were evaluated: Random Forest for machine learning and a Convolutional Neural Network (CNN) for deep learning. Feature extraction focused on texture, morphological characteristics, and image quality. Performance metrics included accuracy, detection rate, and F-measure. The CNN model achieved 92.4% accuracy, outperforming the Random Forest method. The study concluded that the CNN approach was more effective in

diagnosing COVID-19 from X-ray images, emphasizing the importance of automated and accurate detection systems during the COVID-19 pandemic.

Lailis Syafaah et al. [26] used the Hepatitis C Data Set from the UCI Machine Learning Repository for their study. The dataset contains laboratory test results for 73 patients, including 52 males and 21 females, with an age range of 19 to 75 years. It includes 10 parameters: Albumin (ALB), Bilirubin (BIL), Choline esterase (CHE), Gamma-glutamyl transferase (GGT), Aspartate aminotransferase (AST), Alanine aminotransferase (ALT), Cholesterol (CHOL), Creatinine (CREA), Protein (PROT), and Alkaline phosphatase (ALP).The algorithms tested included k-Nearest Neighbors (KNN), Naïve Bayes, Neural Network (NN), and Random Forest (RF). The best performing algorithm was Neural Network (NN), achieving an accuracy of 95.12%.

Muhammad Bilal Butt et al. [27] used the "HCV-Egy-Data" dataset obtained from the UCI Machine Learning Repository for their study. The dataset includes 1385 observations, each with 29 properties, out of which 19 were selected for the study. It contains information on various attributes related to Hepatitis C patients, with the "histological staging" attribute indicating the stage of the patient. The dataset comprises cases distributed across different stages of Hepatitis C.The main algorithm used in the study was the Artificial Back-Propagation Neural Network. Other algorithms used in related studies for predicting the fibrosis stage in Hepatitis C patients included Extreme Gradient Boosting, Support Vector Machine (SVM), Random Forest, K-Nearest Neighbors (KNN), Decision Trees, and Neurofuzzy.The proposed IHSDS algorithm achieved a precision of 98.89%, and during validation, it achieved a precision of 94.44%.

Jagdeep Singh et al. [28] used the ILDP dataset and employed the WEKA tool for feature selection to enhance the performance of machine learning models. They used 10-fold cross-validation to ensure reliable results. The algorithms tested included Logistic Regression, SMO, Random Forest, Naive Bayes, J48, and KNN. The results showed that Logistic Regression with feature selection achieved an accuracy of 74.36%, compared to 72.5% without feature selection on the ILDP dataset.

Ali Farzane1 et al. [29] used gene expression data from 13 GEO data series (. Technologies used include R language for preprocessing and XGBoost for classification. The algorithms employed are XGBoost and SHAP for feature importance. The model achieved an accuracy of 97%, sensitivity of 100%, and specificity of 95%. The results identified a 10-gene set as predictive biomarkers for liver cancer stem cells

Chihua Fanga et al. [30] works on the paper which reviews the latest advancements in digital and intelligent liver surgery, emphasizing the transformative impact of digital imaging technology. Key topics include three-dimensional visualization, 3D printing, virtual reality, molecular fluorescence imaging, artificial intelligence-radiomics, abdominal surgery navigation, and new tumor imaging techniques. These technologies enhance diagnosis, surgical planning, and execution, offering improved surgical outcomes and reduced risks. However, the integration of these technologies into clinical practice presents challenges, including ethical concerns, cost-effectiveness, and the need for high-quality clinical trials to validate their efficacy.

Mehrbakhsh Nilashia et al. [31] on the method for hepatitis disease diagnosis using a combination of machine learning techniques. The method includes Non-

linear Iterative Partial Least Squares (NIPALS) for data dimensionality reduction, Self-Organizing Map (SOM) for clustering, and ensembles of Adaptive Neuro-Fuzzy Inference System (ANFIS) for prediction. Decision trees are used to select the most important features. The proposed method is tested on a real-world dataset and shows improved accuracy compared to previous methods. The study demonstrates that ensemble learning can enhance the performance of hepatitis disease diagnosis systems.

Shimaa M. Abd El-Salam Rabab Salama et al. [32] study examines the use of machine learning to predict esophageal varices in Egyptian chronic hepatitis C patients, aiming to reduce unnecessary endoscopies. It analyzed data from 4,962 patients, using 24 clinical variables. Six algorithms were tested: Neural Networks, Naïve Bayes, Decision Tree, Support Vector Machine, Random Forest, and Bayesian Network. The Bayesian Network performed best, achieving 74.8% AUC and 68.9% accuracy. Key predictors included gender, platelet count, albumin, total bilirubin, baseline PCR, liver and spleen condition, stiffness, and prothrombin concentration.

Elizabeth L. Godfrey et al. [33] works on the decreasing predictive accuracy of the Model for End-Stage Liver Disease (MELD) score over time, particularly in the context of changing liver disease etiologies and demographic shifts. Data from 120,156 liver transplant candidates from 2002-2016 shows that MELD's ability to predict 90-day mortality declined from a concordance statistic of 0.80 in 2003 to 0.70 in 2015. This decline is attributed to the changing nature of liver disease, with conditions like alcoholic liver disease and non-alcoholic fatty liver disease becoming more prevalent.

Amine M. Remita et al. [34] works on the paper "Statistical Linear Models in Virus Genomic Alignment-free Classification: Application to Hepatitis C Viruses" discusses the use of statistical learning methods for classifying viral sequences without relying on sequence alignment. The study explores the effectiveness of generative and discriminative linear classifiers in genotyping and subtyping Hepatitis C Virus (HCV) genomes. Various factors, including classifier types, hyper-parameters, sequence lengths, and k-mer lengths, are evaluated. The research demonstrates that accurate classification can be achieved using k-mer counts from complete genomes, addressing challenges posed by fragmented sequences generated by modern sequencing technologies.

Jacob C et al. [35] works on the application of advanced machine learning techniques to identify undiagnosed patients with Hepatitis C Virus (HCV). By leveraging state-of-the-art methods, the research aims to improve the detection and diagnosis of HCV in patients who might not have been identified through traditional methods. This approach has the potential to enhance early treatment and management of the disease, ultimately leading to better health outcomes for patients.

M. Stepanova et al. [36] study focuses on creating a mapping algorithm for the Chronic Liver Disease Questionnaire-Hepatitis C Version (CLDQ-HCV) to estimate health utility scores from SF-6D, which are derived from the short form 36 health survey. By analyzing data from 34,822 patients with HCV, the researchers developed and tested various regression models. Mixed models achieved a root mean square error (RMSE) of up to 0.088 and Pearson correlations between 0.81 and 0.82. Generalized linear models demonstrated improved accuracy with an RMSE of up to 0.0839 and correlations reaching 0.844. These models maintained

high accuracy across different subpopulations and stages of cirrhosis, showing excellent group-level precision.

J. Cai et al. [37] study presents an automatic diagnosis system for chronic hepatitis C, designed to non-invasively predict the fibrosis stage and inflammatory activity grade using serum indices data. Utilizing an extreme learning machine, which is noted for its straightforward structure and rapid computation, the system demonstrates excellent diagnostic performance. When tested on actual clinical cases, the experimental results indicate that this approach surpasses current state-of-the-art methods in accurately diagnosing the fibrosis stage and inflammatory activity grade in chronic hepatitis C patients, highlighting its efficacy and efficiency in clinical applications.

R. Wei et al. [38] predicting advanced hepatic fibrosis and cirrhosis is difficult due to the invasive nature of liver biopsies. This study investigates the use of less invasive blood tests combined with advanced machine learning algorithms for diagnosis. Machine learning models were developed and compared to the FIB-4 score using a discovery dataset of 490 hepatitis B virus (HBV) patients, then validated with an independent dataset of 86 HBV patients. Furthermore, the models were applied to two independent hepatitis C virus (HCV) datasets, consisting of 254 and 230 patients, to evaluate their applicability. The results highlight the potential of these machine learning approaches in accurately diagnosing advanced hepatic fibrosis and cirrhosis.

S. Hashem et al. [39], evaluates machine learning techniques for predicting advanced liver fibrosis in chronic hepatitis C patients, aiming to replace invasive biopsies. Using a cohort of 39,567 patients categorized by METAVIR scores (F0-F2 and F3-F4), the study developed decision tree, genetic algorithm, particle

swarm optimization, and multi-linear regression models. Significant predictors included age, platelet count, AST, and albumin. The models achieved AUROC values of 0.73 to 0.76 and accuracy between 66.3% and 84.4%. The findings suggest that machine learning approaches can effectively predict advanced fibrosis, offering a non-invasive alternative to traditional methods.

Naga Chalasani's et al. [40] guidance on NAFLD diagnosis and management emphasized evaluating patients based on clinical and imaging findings rather than a specific dataset. The document recommends several noninvasive assessment tools, including Clinical Decision Aids (NAFLD Fibrosis Score, FIB-4 index, APRI), serum biomarkers (ELF panel, Fibrometer, FibroTest, Hepascore), and imaging techniques (Transient Elastography, Magnetic Resonance Elastography, Acoustic Radiation Force Impulse Imaging). Management should address liver disease and metabolic comorbidities (obesity, insulin resistance, T2DM), with pharmacological treatments for biopsy-proven NASH and fibrosis. Systematic screening is crucial, especially in high-risk populations, as normal liver biochemistries can complicate diagnosis. The guidance underscores the importance of combining clinical assessment and noninvasive tools for effective NAFLD management and early intervention.

Emmanuel Agosto-Arroyo et al. [41] works on the development and implementation of Alchemy, a web-based, real-time quality assurance (QA) platform designed for the molecular diagnostics laboratory. The platform aims to improve test turnaround time (TAT) for human immunodeficiency virus (HIV), hepatitis C virus (HCV), and BK virus (BKV) quantitation assays. Using a Linux, Nginx, MariaDB, PHP stack, Alchemy automates QA reporting by processing data from laboratory information systems (LISs) and generating reports. This

automation reduces the time required for QA report preparation from 45-60 minutes per test to 15 minutes per month, minimizing human error and enhancing efficiency. The platform's development showcases the benefits of integrating informatics with clinical workflows.

Rohit Loomba et al. [42] presents a metagenomic signature based on gut microbiome data for the non-invasive detection of advanced fibrosis in patients with non-alcoholic fatty liver disease (NAFLD). The study highlights the association between gut microbiota and advanced fibrosis, demonstrating that a gut microbiome panel can accurately diagnose advanced fibrosis. The findings suggest a potential non-invasive diagnostic tool for NAFLD-related advanced fibrosis.

Yoichi Hayashi et al. [43] investigates the accuracy and interpretability of diagnostic rules for liver disease using a new rule extraction algorithm, Continuous Re-RX combined with sampling selection techniques (Sampling-Continuous Re-RX). This algorithm aims to produce highly accurate and interpretable rules by balancing the trade-off between accuracy and the number of rules extracted. The performance was evaluated using the BUPA and Hepatitis datasets, and compared with previous techniques, showing improved accuracy and interpretability. The findings highlight the algorithm's potential in medical settings for diagnosing liver disease based on serum biomarkers and Child-Pugh scores.

Tapas Ranjan Baitharu et al. [44] analyzes various data mining techniques for developing a healthcare decision support system using a liver disorder dataset. It aims to improve early detection of liver diseases like cirrhosis and hepatitis by comparing classifiers such as J48, Naive Bayes, ANN, ZeroR, 1BK, and VFI. The

study finds that accurate data classification is crucial for effective medical diagnosis, offering insights into the predictive performances of these classifiers, and emphasizes the importance of data mining in healthcare for better disease prediction and decision-making.

Ki Tae Suk et al. [45] highlighted the importance of HVPG measurement in staging liver fibrosis or cirrhosis. They found HVPG measurement to be crucial for evaluating portal pressure with an AUROC of 0.85 for predicting advanced fibrosis in chronic viral hepatitis, showing 80% sensitivity and 77% specificity, outperforming serologic biomarkers. Their review noted that liver biopsy, although a gold standard, may have sampling errors. HVPG measurement provides substantial prognostic information post-treatment, but combining HVPG with MELD/MELD-Na scores does not improve prognostic accuracy. The study was supported by grants from the National Research Foundation of Korea and the Rural Development Administration.

Nahum Mendez-Sanchez et al. [46] titled "Latin American Association for the Study of the Liver Recommendations on Treatment of Hepatitis C" addresses the major public health concern posed by chronic hepatitis C virus (HCV) infection. It estimates that over 185 million people globally live with chronic hepatitis C, with significant risk of cirrhosis and hepatocellular carcinoma. The guidelines cover various treatment options, including direct-acting antivirals (DAAs) and their efficacy across different HCV genotypes. Public health policies, diagnostic methods, and the cost of treatment in Latin American countries are also discussed. The importance of surveillance and accurate data collection for effective policy-making is emphasized.

KayvanJoo et al. [47] aimed to predict the outcome of hepatitis C virus (HCV) therapy using machine learning algorithms. They analyzed full-length nucleotide sequences of HCV subtypes 1a and 1b to identify genetic markers linked to therapy response. The research utilized various feature selection methods and machine learning techniques, achieving prediction accuracies of up to 85% for distinguishing between responders and non-responders. The study found specific nucleotide attributes, such as counts of certain base pairs, which were crucial in predicting therapy outcomes. These findings could guide personalized HCV treatments and enhance therapeutic success rates.

Mahmoud ElHefnawi et al. [48] used consists of clinical and biochemical data from 200 Egyptian patients with Hepatitis C, treated with Pegylated-Interferon and Ribavirin. Machine learning techniques employed include Artificial Neural Networks (ANN) and Decision Trees (DT). ANN architectures ranged from 70 to 180 neurons in the hidden layer, while DT utilized the CART classification algorithm. The best accuracy achieved was 0.80 for DT and 0.76 for ANN. ANN sensitivity and specificity were 0.95 and 0.39 respectively, while DT achieved 0.89 and 0.78 respectively. The study concluded that DT provided higher predictive accuracy.

Weifeng Shi1 et al. [49] "Recombination in Hepatitis C Virus: Identification of Four Novel Naturally Occurring Inter-Subtype Recombinants" by Courtney S. F. Atkinson, Vincent C. Emery, and Judith Breuer analyzed 1,278 full-length HCV genome sequences from the Los Alamos HCV database to investigate recombination events. Using Clustal Omega for protein sequence alignment, Bioedit for DNA alignment, and RAxML for phylogenetic tree construction, the

study identified four novel inter-subtype recombinants (6a/6o, 6e/6h, 6e/6o, 6n/6o) with high accuracy, indicated by very low P values (<10^−20, <10^−10, <10^−5). This discovery suggests that HCV recombination, while rare, is significant and necessitates further research to understand its impact on HCV pathogenesis and clinical outcomes.

E.F.Duffell et al. [50] study revealed significant discrepancies in reported cases among countries, reflecting variations in testing practices and disease awareness. Most hepatitis C cases were reported among young adult males, with injecting drug use being the primary transmission route. Despite improvements in data completeness, many cases remained classified as 'unknown.' Enhanced surveillance and better understanding of testing practices are needed to accurately assess the true burden of hepatitis C in Europe and improve public health strategies.

2.1 Research Gaps

Research gaps in liver disease prediction using machine learning can stem from several factors:

1. **Data Limitations:**
    - **Data Quality and Volume:** Inadequate or poor-quality data can hinder the performance of machine learning models. Incomplete or inconsistent records may affect the accuracy and reliability of predictions.
    - **Data Biases:** If the dataset does not accurately represent the broader patient population, it may result in biased models that do not generalize well to diverse groups.
2. **Algorithmic Limitations:**

- **Complexity of Models:** Advanced models like deep learning or ensemble techniques, while potentially highly effective, often demand significant computational resources and can be difficult to interpret.
- **Feature Selection Issues:** Poor feature selection can lead to overfitting or underfitting, which affects model performance. Effective extraction and selection of features are critical for creating accurate predictive models.

3. **Validation and Generalization:**

- **External Validation Challenges:** Models developed on specific datasets may not perform well when applied to other datasets or populations. It's essential to validate models on external datasets to ensure their robustness.
- **Reproducibility Concerns:** Differences in research methods, data preprocessing, and evaluation metrics can hinder the reproducibility of results and complicate comparisons across studies.

4. **Clinical Relevance:**

- **Integration into Clinical Practice:** Adapting machine learning models for clinical use involves addressing practical issues such as user-friendliness, integration with existing systems, and training for healthcare professionals.
- **Consistency in Outcome Measures:** Varying outcome measures across studies can affect the comparability of results.

Consistent and relevant measures are necessary for accurate evaluation of model performance.

5. **Technological Constraints:**

    - **Resource Requirements:** Advanced machine learning approaches may require substantial computational resources, which may not be available in all research environments.
    - **Data Privacy and Security:** Managing sensitive patient data involves privacy and security concerns that can limit data sharing and collaborative research efforts.

6. **Ethical and Regulatory Considerations:**

    - **Ethical Issues:** The application of machine learning in healthcare must address ethical concerns, including fairness, transparency, and accountability in decision-making processes.
    - **Regulatory Compliance:** Ensuring compliance with regulations and standards for medical data and algorithmic decision-making is crucial, though it can also present challenges to research and implementation.

Addressing these gaps requires enhancing data quality and quantity, advancing algorithm development, ensuring thorough validation, and considering practical and ethical aspects of implementing machine learning models in healthcare.

CHAPTER-3

PROPOSED METHODOLOGY

3.1 Proposed System

The system will be continuously improved through the integration of advanced machine learning techniques in future iterations. The process begins with data visualization to explore and understand the HCV dataset comprehensively. This is followed by a robust data pre-processing stage, which includes data cleaning, normalization, and attribute verification.

After pre-processing, the data is divided into training and testing sets. To address any potential data imbalance, the SMOTE technique is applied as needed. The next step involves creating the model by selecting suitable algorithms and optimizing hyperparameters using Optuna.

The final stages involve thorough model evaluation to ensure performance accuracy, culminating in reliable prediction outcomes. To maintain and enhance the system's accuracy and adaptability, continuous model retraining and updating are essential.

Here is a revised version of the flowchart methodology for creating a machine learning model to predict liver disease:

3.2 Methodology

Step 1. **Data Collection**: Compile patient data, including various blood test results and demographic information.

Step 2. **Data Preprocessing**: Cleanse and transform the data, address any missing values, and normalize the features.

Step 3. **Feature Selection**: Identify and select the most relevant features for model training.

Step 4. **Model Selection**: Determine the most suitable machine learning algorithms for the task.

Step 5. **Model Training**: Train the models using the training dataset.

Step 6. **Hyperparameter Tuning**: Refine the model's hyperparameters with Optuna.

Step 7. **Model Evaluation**: Assess the models using both validation and test datasets.

Step 8. **SMOTE Analysis**: Use SMOTE to address class imbalances in the dataset.

Step 9. **Model Interpretation**: Apply SHAP to interpret the model's predictions.

Step 10. **Deployment**: Deploy the model that demonstrates the best performance.

3.2.1 Flowchart

The flowchart for the proposed model is shown in figure 2.1

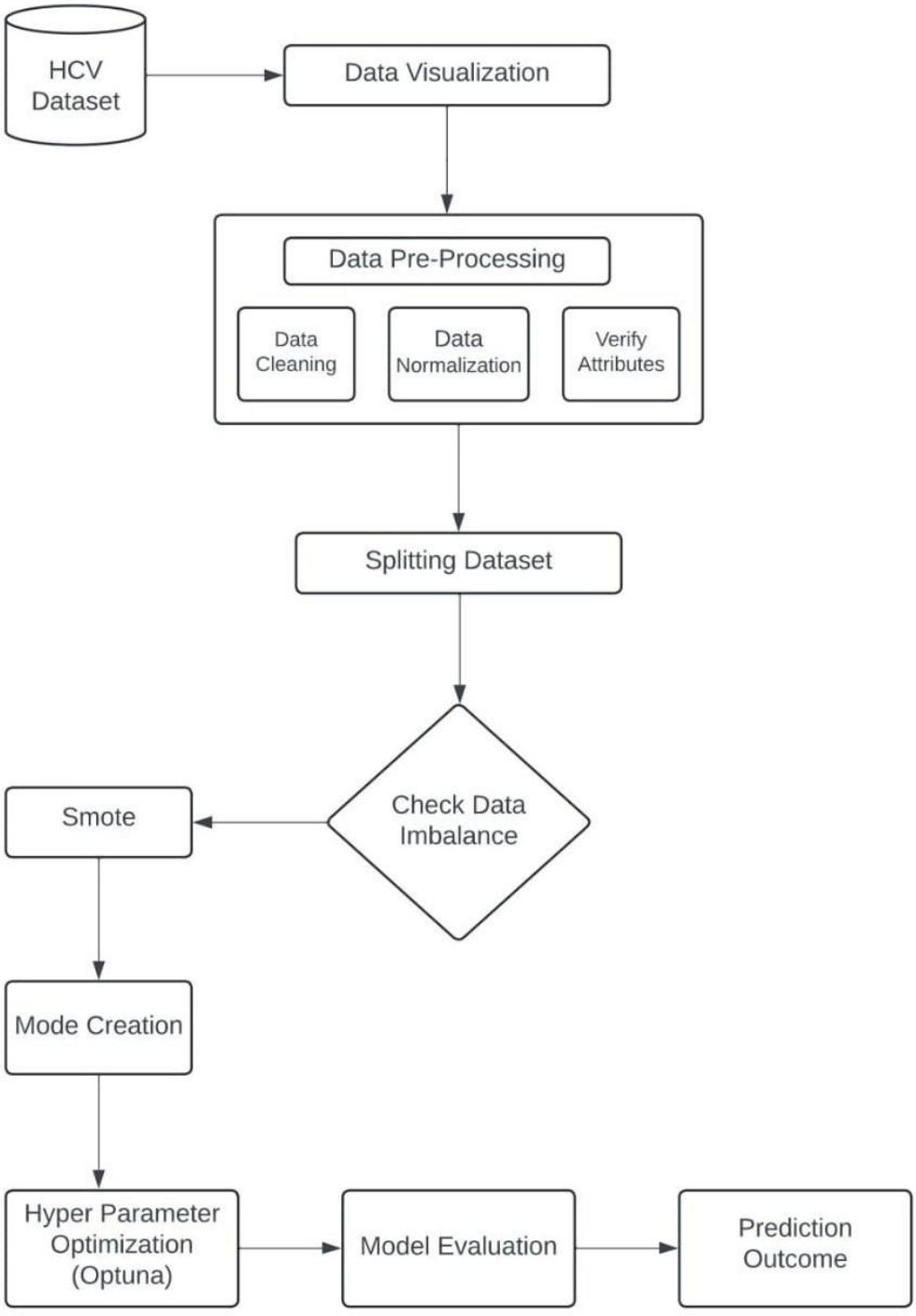

Figure: 3.1 Flowchart of Proposed Model

3.3 MATERIALS AND METHODS

3.3.1 Dataset

(615 rows, 13 columns)

List Of Attributes:

1. **Category**: The target variable. Values include: '0 = Blood Donor', '0s = Suspect Blood Donor', '1 = Hepatitis', '2 = Fibrosis', '3 = Cirrhosis'.
2. **Age**: The patient's age in years.
3. **Sex**: The patient's gender ('f' = female, 'm' = male).
4. **ALB**: Concentration of albumin in the patient's blood.
5. **ALP**: Level of alkaline phosphatase in the patient's blood.
6. **ALT**: Concentration of alanine transaminase in the patient's blood.
7. **AST**: Amount of aspartate aminotransferase in the patient's blood.
8. **BIL**: Level of bilirubin in the patient's blood.
9. **CHE**: Concentration of cholinesterase in the patient's blood.
10. **CHOL**: Level of cholesterol in the patient's blood.
11. **CREA**: Amount of creatinine in the patient's blood.
12. **GGT:** Concentration of gamma-glutamyl transferase in the patient's blood.
13. **PROT:** Amount of protein in the patient's blood.

3.3.2 Libraries and Modules

1. **NumPy**

- **Purpose**: Executes linear algebra operations.

- **Key Functionality**: NumPy is vital for scientific computing in Python, offering support for arrays, matrices, and a range of mathematical functions.

2. **Pandas**

- **Purpose**: Assists with data manipulation and analysis.
- **Key Functionality**: Pandas provides powerful data structures, such as DataFrames, which are essential for data cleaning, transformation, and analysis.

3. **Matplotlib**

- **Purpose**: Generates data visualizations.
- **Key Functionality**: Matplotlib is a flexible plotting library used for creating static, animated, and interactive visualizations.

4. **Scikit-learn**

- **Purpose**: Facilitates machine learning, data preprocessing, model creation, and evaluation.
- **Key Functionality**: Scikit-learn is a comprehensive machine learning library that offers tools for preprocessing, model selection, and evaluation, supporting a broad range of machine learning algorithms and utilities.

Key Components:

- **ColumnTransformer**: Applies distinct preprocessing techniques to different subsets of features.
- **StandardScaler**: Normalizes features by removing the mean and scaling them to unit variance.

- **KNNImputer**: Imputes missing values using the k-nearest neighbors method.
- **Pipeline**: Links multiple transformers and estimators to simplify the machine learning workflow.
- **PCA**: Principal Component Analysis for reducing dimensionality.
- **Train_test_split**: Divides the dataset into training and testing subsets.
- **RandomForestClassifier**: Improves performance by combining multiple decision trees.
- **VotingClassifier**: Enhances predictive performance by aggregating the predictions of several models.
- **LogisticRegression**: A linear model designed for binary classification.
- **KNeighborsClassifier**: Implements the k-nearest neighbors algorithm for classification.
- **SVC**: Support Vector Classifier for classification tasks.
- **GaussianNB**: Gaussian Naive Bayes classifier for classification.

5. **Imbalanced-learn**

- **Purpose**: Addresses imbalanced datasets.
- **Key Functionality**: Imbalanced-learn offers methods for managing class imbalances in datasets, including SMOTE (Synthetic Minority Over-sampling Technique), which creates synthetic samples to balance class distributions.

6. **Optuna**

- **Purpose**: Optimizes hyperparameters.
- **Key Functionality**: Optuna is a powerful framework for automated hyperparameter tuning, designed to enhance model performance by discovering optimal hyperparameters.

3.3.3 Explanation of the Algorithms and Their Appropriateness

- **Logistic Regression:**
  - Ideal for binary classification tasks.
  - Offers probability estimates, which can be valuable for medical diagnosis.
  - Regularization methods (L1, L2, Elastic Net) assist in managing multicollinearity and feature selection.
- **K-Nearest Neighbors (KNN):**
  - A straightforward and easy-to-understand algorithm.
  - Particularly effective for smaller datasets or those with simple decision boundaries.
  - Hyperparameters such as the number of neighbors, distance metric, and weighting function can be adjusted.
- **Support Vector Machine (SVM):**
  - Effective for high-dimensional spaces.
  - Suitable for problems where the decision boundary is not linear.
  - Kernel functions (such as polynomial and RBF) can be fine-tuned to enhance performance.
- **Random Forest:**

- An ensemble technique that aggregates multiple decision trees.
- Offers strong performance with a lower risk of overfitting.
- Well-suited for managing complex data patterns and interactions between features.

- **Naive Bayes (GaussianNB):**
  - Simple and computationally efficient.
  - Assumes feature independence, which may not always be accurate but can still be effective for certain types of data.
  - Useful for initial benchmarking and handling high-dimensional data.

By optimizing these algorithms and combining their strengths through a Voting Classifier, we can enhance their collective predictive capabilities to achieve greater accuracy on the dataset.

3.3.4 Data Visualization:

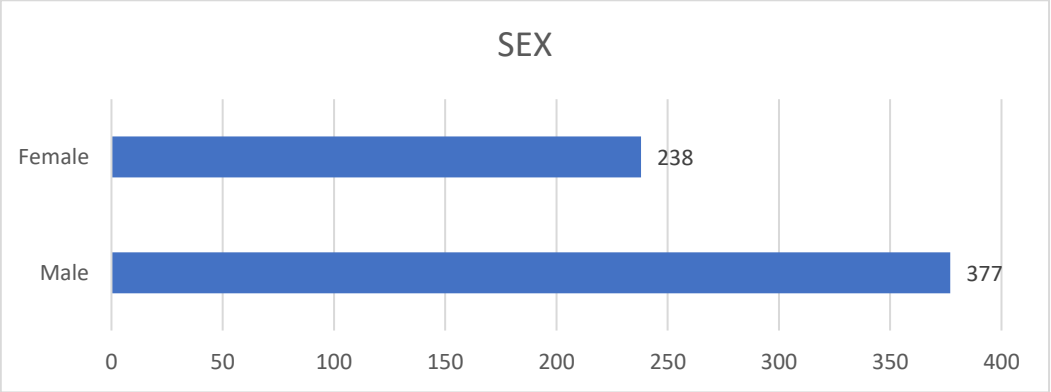

Figure: 3.2 Sex Ratio Graph

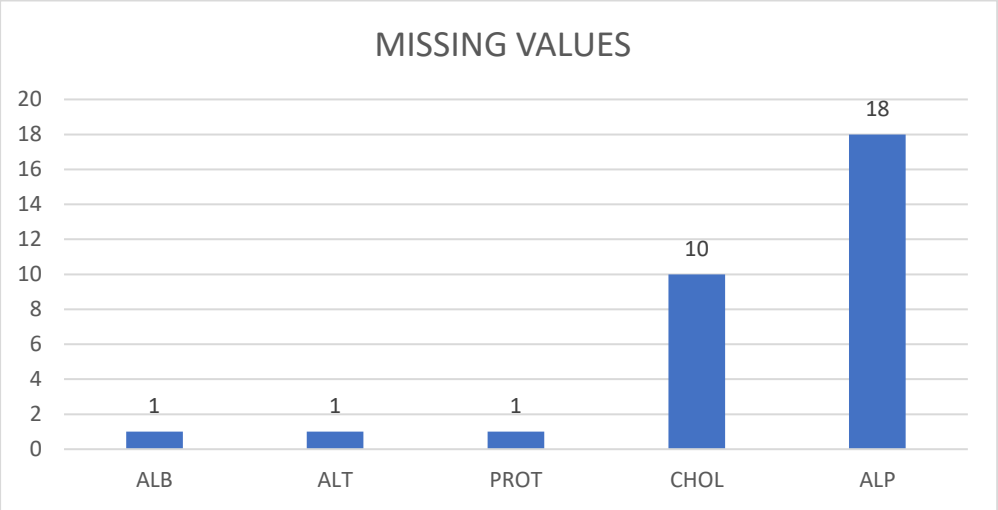

Figure: 3.3 Missing Values

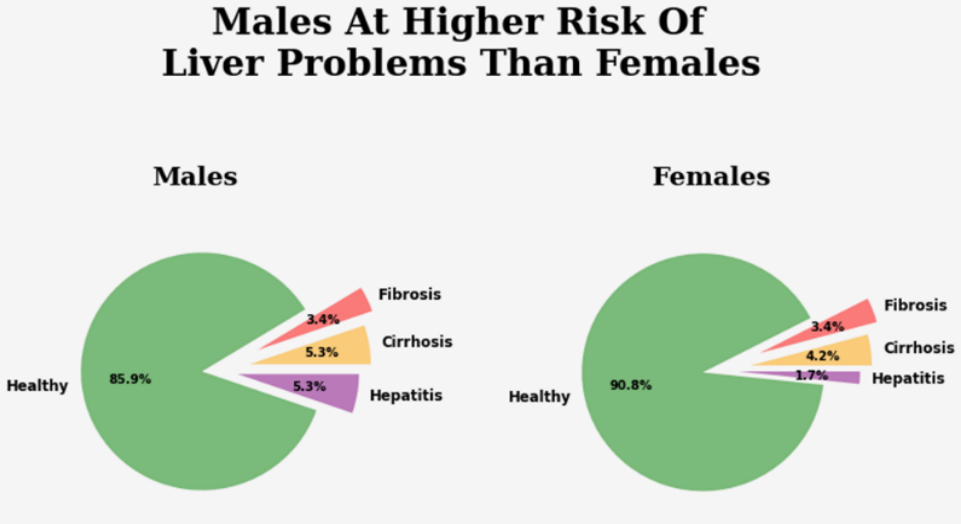

Figure: 3.4 Risk Comparison

3.3.5 SMOTE Analysis

Introduction to SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) is a prevalent method for addressing class imbalance in datasets. Class imbalance occurs when some classes are significantly underrepresented compared to others, leading to biased model outcomes. SMOTE combats this issue by generating synthetic samples for the minority class, thereby balancing the dataset and improving the model's ability to generalize across all classes.

How SMOTE Operates

SMOTE creates new synthetic instances for the minority class through interpolation between existing minority class samples. The process involves the following steps:

1. Select Minority Class Instances: Randomly choose an instance from the minority class.
2. Identify Nearest Neighbors: Find the k-nearest neighbors (typically k=5) of the selected instance within the minority class.
3. Create Synthetic Instances: Generate synthetic samples by interpolating new instances along the line segments connecting the selected instance with its nearest neighbors.

Advantages of Using SMOTE

- Improves Model Performance: Balancing the class distribution enables the model to better recognize and learn from minority class instances.
- Reduces Bias: It addresses the bias towards the majority class, leading to enhanced performance across all classes.
- Boosts Generalization: A more balanced dataset improves the model's capacity to generalize to new, unseen data, thus minimizing the risk of overfitting.

3.3.6 Dataset Splitting

The dataset is divided into three separate subsets: training, validation, and test sets. This separation ensures accurate model evaluation and helps prevent data leakage.

- Training Set: Used to train the model.
- Validation Set: Utilized to fine-tune model parameters and validate performance during the training phase.
- Test Set: Employed to evaluate the performance of the final model.

3.3.7 Optuna Hyperparameter Optimization

Optuna is an open-source tool designed to automate hyperparameter tuning for machine learning models. Renowned for its efficiency, flexibility, and ease of use, Optuna offers several key features, including automatic search space exploration, trial pruning, and result visualization tools.

Reasons to Use Optuna

- **Efficiency**: Optuna utilizes advanced optimization algorithms like the Tree-structured Parzen Estimator (TPE) to effectively find optimal hyperparameters, reducing both time and computational resources compared to traditional methods such as grid or random search.
- **Flexibility**: It supports a wide range of machine learning frameworks and allows for the definition of complex search spaces.
- **Automation**: Optuna automates the hyperparameter tuning process, allowing developers to focus more on model development rather than manual tuning.

- **Pruning**: Optuna can terminate less promising trials early, conserving time and computational resources by avoiding evaluations of configurations that are unlikely to perform well.

Application of Optuna in This Project

In this project, Optuna is employed to improve the performance of a predictive model for diagnosing liver diseases. The process involves several key steps:

- **Defining the Objective Function**: The function that Optuna optimizes includes:
    - Splitting the dataset into training and validation sets.
    - Training the machine learning model (e.g., Random Forest, SVM) with specific hyperparameters.
    - Evaluating the model's performance on the validation set using metrics such as accuracy or F1 score.
- **Setting Up the Search Space**: This defines the range of possible values for each hyperparameter. For example:
    - For a Random Forest model: parameters like the number of trees, maximum depth, and minimum samples split.
    - For an SVM model: parameters such as kernel type, regularization parameter (C), and gamma.
- **Running the Optimization**: Optuna performs multiple trials to explore the search space and identify the best hyperparameters.

- **Pruning**: During the optimization process, Optuna prunes trials that show poor performance based on intermediate metrics, thereby concentrating resources on more promising configurations.
- **Analyzing the Results**: After the optimization, Optuna provides tools for visualizing and analyzing results to understand the impact of different hyperparameters on model performance.

By using Optuna for hyperparameter optimization, the model's performance is significantly enhanced through systematic exploration and tuning of hyperparameters. This results in a more robust and accurate predictive model for diagnosing liver diseases, as indicated by high validation and test accuracies. Optuna ensures the optimal hyperparameter configuration is applied, thereby maximizing the model's predictive capability and reliability.

3.3.8 SHAP in Machine Learning

SHAP (SHapley Additive exPlanations) is a method for explaining the predictions of machine learning models. It offers a standardized approach to interpret complex models by assigning an importance value to each feature for a given prediction. SHAP values are rooted in cooperative game theory, specifically the Shapley value, which provides a fair distribution of the total prediction among the features.

Reasons to Use SHAP

- **Interpretability**: SHAP values deliver insights into individual predictions, clarifying the rationale behind a model's decisions. This transparency is crucial for fostering trust in models, particularly in high-stakes fields like healthcare, finance, and law.

- **Global and Local Explanations**: SHAP offers both global insights (overall feature importance across the dataset) and local insights (feature contributions for specific predictions).
- **Model-Agnostic**: SHAP can be applied to any type of machine learning model, ranging from simple linear regressions to complex ensemble methods such as gradient boosting and random forests.
- **Consistency and Fairness**: SHAP ensures consistency and additivity. A feature's SHAP value increases with its contribution to the model's prediction, and the sum of SHAP values equals the difference between the model's prediction and the baseline prediction.
- **Visualization**: SHAP provides various visualization tools, including summary plots, dependence plots, and force plots, which help illustrate feature impacts on predictions and make model behavior more accessible to non-technical audiences.

How SHAP Functions

- **Baseline Prediction**: SHAP values describe the deviation of a model's prediction for a specific instance from the average prediction across the dataset (the baseline).
- **Coalition Formation**: SHAP evaluates all possible feature subsets (coalitions) to determine each feature's marginal contribution to the prediction.
- **Shapley Value Calculation**: The feature contributions are averaged across all potential coalitions to ensure a fair distribution of the prediction

difference among features. This involves calculating the marginal contribution of each feature by including it in all possible combinations with other features and averaging these contributions.

SHAP is an effective tool for interpreting machine learning models, offering both global and local insights into predictions. Its ability to maintain consistency, fairness, and model-agnostic interpretability makes it valuable for explaining complex models and building trust in their outputs.

3.4 EVALUATION METRICS

Evaluation metrics in machine learning are essential tools for assessing a model's performance. They help determine the accuracy and effectiveness of a model's predictions and whether it achieves the intended goals. The selection of an appropriate metric depends on the specific problem being addressed, whether it is classification, regression, or clustering.

3.4.1 Accuracy

Accuracy is a metric in machine learning used to gauge the performance of a classification model. It is calculated as the proportion of correctly predicted instances out of the total instances.

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

Why we use accuracy?

Accuracy is employed as a performance metric in machine learning for several reasons:

1. Simplicity

2. Interpretability
3. Baseline Performance
4. Initial Assessment

Despite its utility, accuracy has limitations, particularly with imbalanced datasets where the majority class skews the metric. In such cases, alternative metrics like precision, recall, F1 score, and ROC-AUC might offer a more comprehensive evaluation of model performance.

When to use accuracy

Accuracy is most suitable as a performance metric in the following situations:

1. Balanced Datasets
2. Equal Importance of Classes
3. Initial Model Evaluation
4. Baseline Comparison

3.4.2 Precision

Precision is a performance metric in classification problems that measures the accuracy of the model's positive predictions. It is calculated as the ratio of true positive predictions to the total number of positive predictions (true positives and false positives).

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

Why to use Precision:

Precision is utilized as a performance metric in machine learning for several key reasons:

1. Focus on Positive Predictions
2. Handling Imbalanced Datasets
3. Minimizing False Positives
4. Balanced with Recall
5. Model Evaluation and Tuning

Why to use Precision

Precision is most appropriate in the following scenarios:

1. Imbalanced Datasets
2. High Cost of False Positives
3. Binary Classification

3.4.3 Recall

Recall, also known as sensitivity, is a performance metric used in classification problems to measure a model's ability to identify all relevant instances. It is defined as the ratio of true positive predictions to the total number of actual positive instances.

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

Why to use Recall

Recall is an important performance metric in classification problems for several reasons:

1. High Cost of Missing Positive Cases:
    - Critical Applications
    - Safety and Compliance.
2. Imbalanced Datasets:
    - Minority Class Focus
3. Comprehensive Detection:
    - Broad Coverage
4. Initial Screening:
    - Sensitivity Prioritization
5. Model Evaluation and Comparison:
    - Holistic Performance Evaluation

When to use Recall

Recall is particularly useful in the following situations:

1. High Cost of Missing Positive Cases
2. Imbalanced Datasets
3. Critical Applications

3.4.4 F1 score

The **F1 Score** is a performance metric used in classification problems to strike a balance between precision and recall. It is the harmonic mean of precision and recall, offering a single metric that takes into account both false positives and false negatives. This metric is particularly beneficial when it is necessary to balance precision and recall, especially in cases of imbalanced datasets.

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

Why Use the F1 Score?

The F1 score is valuable for several reasons:

1. Balancing Precision and Recall
    - Trade-off Management
    - No Single Dominant Metric
2. Handling Imbalanced Datasets
    - Reflecting True Performance
3. Critical Applications
    - Applications Requiring Balanced Performance
4. Model Comparison and Optimization
    - Unified Metric
    - Optimization Objective

CHAPTER-4

RESULTS AND DISCUSSION

4.1 Results Obtained

Using Voting Classifiers

General Model:

Metric	**Accuracy**
Validation Accuracy	**0.96**
Test Accuracy	**0.95**

Table: 4.1 Accuracy of General Model

SVM MODEL

Metric	**Accuracy**
Best Hyperparameters	0.9354838

Table: 4.2 Accuracy of SVM Model

Decision Tree:

Metric	**Value**
Accuracy	0.946236559139785
Precision	0.953405017921147
Recall	0.946236559139785
F1 Score	0.9483870967741935

Table: 4.3 Evaluation Metrics of Decision Tree

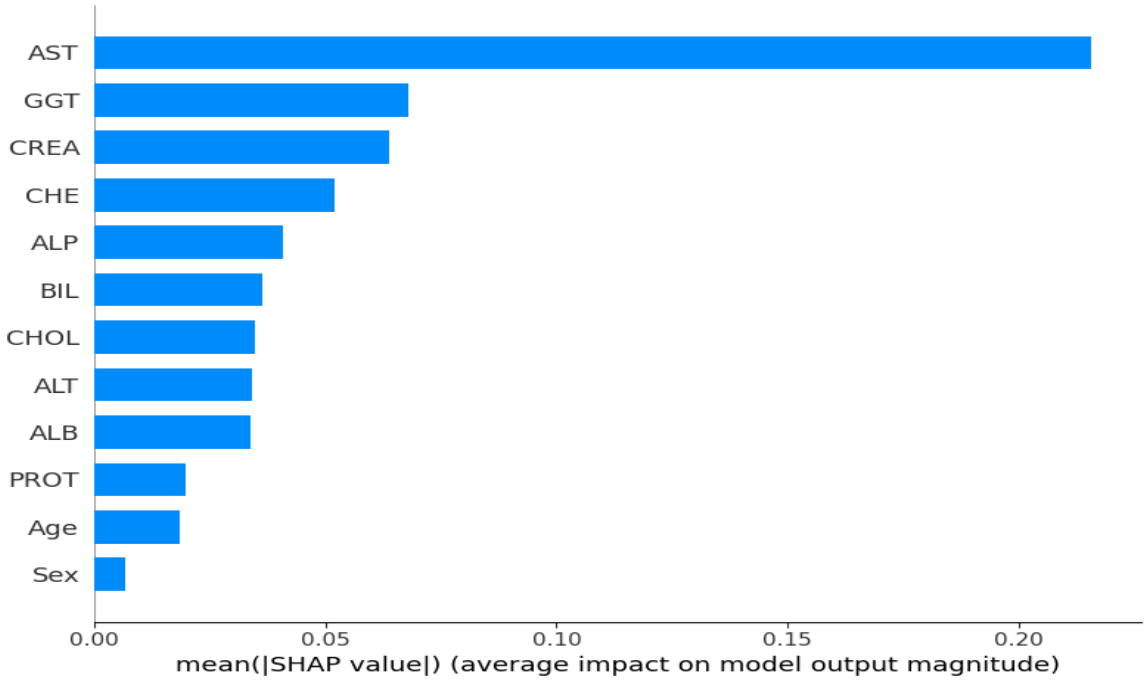
AST
GGT
CREA
CHE
ALP
BIL
CHOL
ALT
ALB
PROT
Age
Sex
0.00
0.05
0.10
0.15
0.20
mean(|SHAP value|) (average impact on model output magnitude)

Fig 4.1

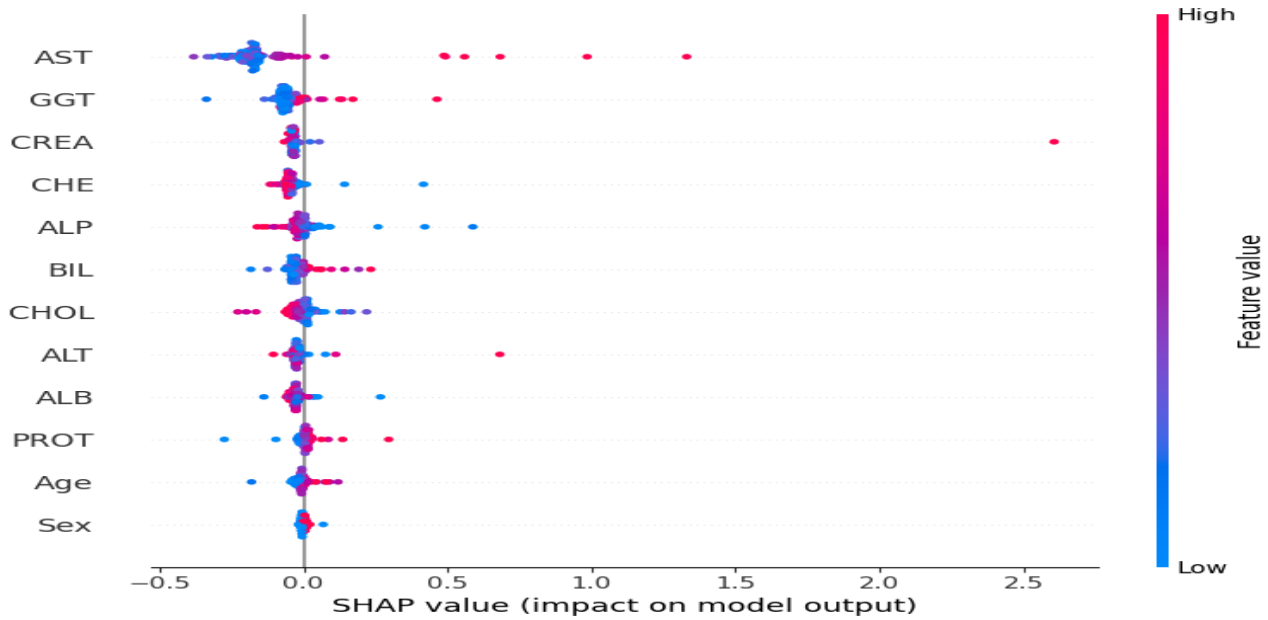
AST
GGT
CREA
CHE
ALP
BIL
CHOL
ALT
ALB
PROT
Age
Sex
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
SHAP value (impact on model output)
High
Feature value
Low

Fig 4.2

4.2 Comparative Analysis

The comparative analysis of various techniques across different studies shows a range of outcomes. Random Forest (RF) consistently performs well across multiple studies, achieving high accuracy rates (89% by N.K. Trivedi et al. and 89.6% by N. Chalasani et al.). Gradient Boosting (GB) achieved the highest individual model performance with 91.29% accuracy as reported by D. Omran et al.

However, the proposed model, which uses a voting classifier combining multiple techniques, outperforms all individual methods with a 95% accuracy. This suggests that ensemble methods, specifically voting classifiers, may provide significant improvements over single algorithm approaches by leveraging the strengths of multiple models.

REFERENCES	**USED TECHNIQUES**	**OUTCOMES**
N.K.Trivedi et al. [25]	RF, SVM, GB	RF-89%
K.T.Suk et al. [45]	RF, KNN	RF-6% more than KNN
D.Omran et al. [35]	SVM, DT, GB, LR, NB, KNN, XGB, RF	DB-91.29%
N.Chalasani et al. [40]	RF	89.6%
Proposed model	Voting classifier	95%

Table 4.4 Comparative analysis of existing work on hepatitis C prediction

4.3 Discussions

This project focuses on developing a predictive model for diagnosing liver diseases using blood test parameters. By utilizing libraries such as NumPy, Pandas, Matplotlib, Scikit-learn, Imbalanced-learn, and Optuna, the project emphasizes data manipulation, visualization, machine learning, and hyperparameter optimization.

Dataset and Features

The dataset comprises 13 attributes related to blood test results and demographics. The target feature is 'Category,' which identifies the type of liver condition. Key attributes include levels of albumin, alkaline phosphatase, alanine transaminase, aspartate aminotransferase, bilirubin, cholinesterase, cholesterol, creatine, gamma-glutamyl transferase, protein, age, and sex.

Algorithms and Their Suitability

- **Logistic Regression**: Suitable for binary classification, providing probability estimates.
- **K-Nearest Neighbors (KNN)**: Simple and effective for small datasets.
- **Support Vector Machine (SVM)**: Ideal for high-dimensional, non-linear data.
- **Random Forest**: Delivers robust performance by combining multiple decision trees.
- **Naive Bayes (GaussianNB)**: Computationally efficient, though it assumes feature independence.

A VotingClassifier combines these algorithms to enhance predictive accuracy.

SMOTE Analysis

SMOTE (Synthetic Minority Over-sampling Technique) addresses class imbalance by generating synthetic samples for minority classes, thereby improving model performance and reducing bias.

Data Splitting

The dataset is divided into training, validation, and test sets to ensure proper evaluation and prevent data leakage.

Model Performance

- **Validation Set Accuracy**: 0.96
- **Test Set Accuracy**: 0.95

The ensemble model, which combines logistic regression, KNN, SVM, random forest, and Naive Bayes, achieves high accuracy.

CHAPTER-5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

The concentration of certain proteins and amino acids in a patient's blood serves as a strong indicator of their risk for liver disease, especially when considering specific combinations of these markers. Among these, the level of aspartate aminotransferase (AST) in the blood significantly influences the ensemble model's predictions. As illustrated by the SHAP graphs, higher AST levels correlate with an increased likelihood of liver disease classification.

The final model achieves approximately 95% accuracy on test data. This model is an ensemble consisting of five algorithms: logistic regression, k-nearest neighbors, support vector machine, random forest, and naive Bayes. Within this ensemble, random forest and logistic regression are the most influential, predominantly determining the predictions for the HCV class type.

5.2 Future Work

1. **Integration with Clinical Systems**
    - Integrate the predictive model into hospital information systems for real-time liver disease screening.
    - Enable seamless incorporation with electronic health records (EHRs) for comprehensive patient monitoring.
2. **Expansion of Dataset**

- Utilize larger and more diverse datasets from various demographics to enhance model generalizability.
- Include additional biomarkers and clinical data to improve predictive accuracy.

3. **Advanced Model Techniques**
   - Implement deep learning models, such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs), for enhanced feature extraction and prediction.
   - Explore transfer learning techniques to leverage pre-trained models on similar medical datasets.
4. **Real-time Monitoring**
   - Develop a system for continuous monitoring of liver health using regular blood tests and wearable devices.
   - Provide alerts and recommendations based on real-time data analysis.
5. **Patient-Specific Customization**
   - Tailor the predictive model to individual patient profiles for personalized medicine.
   - Incorporate genetic and lifestyle factors to refine risk assessments.

6.Collaboration and Data Sharing

- Facilitate collaboration with research institutions for data sharing and model validation.
- Contribute to open-source projects and public health initiatives to combat liver disease.

6. **Regulatory Approval and Clinical Trials**

- Pursue regulatory approvals for clinical use of the predictive model.
- Conduct extensive clinical trials to validate the model's effectiveness and reliability in diverse clinical settings.

7. **Educational Tools**

- Develop educational resources and tools for healthcare professionals to understand and utilize the model.

REFERENCES

[1] S. Sarkar *et al.*, "A machine learning model to predict risk for hepatocellular carcinoma in patients with metabolic dysfunction-associated steatotic liver disease," *Gastro Hep Adv.*, vol. 3, no. 4, pp. 498–505, 2024.

[2] J. L. Calleja-Panero *et al.*, "Chronic liver disease-associated severe thrombocytopenia in Spain: Results from a retrospective study using machine learning and natural language processing," *Gastroenterol. Hepatol. (Engl. Ed.)*, vol. 47, no. 3, pp. 236–245, 2024.

[3] Q. Chen *et al.*, "Personalized prediction of postoperative complication and survival among Colorectal Liver Metastases Patients Receiving Simultaneous Resection using machine learning approaches: A multi-center study," *Cancer Lett.*, vol. 593, no. 216967, p. 216967, 2024.

[4] B. Wang *et al.*, "Machine learning deciphers the significance of mitochondrial regulators on the diagnosis and subtype classification in non-alcoholic fatty liver disease," *Heliyon*, vol. 10, no. 9, p. e29860, 2024.

[5] E. Ruiz *et al.*, "A preoperative risk score based on early recurrence for estimating outcomes after resection of hepatocellular carcinoma in the non-cirrhotic liver," *HPB (Oxford)*, vol. 26, no. 5, pp. 691–702, 2024.

[6] A. M. Elshewey *et al.*, "Optimizing HCV disease prediction in Egypt: The hyOPTGB framework," *Diagnostics (Basel)*, vol. 13, no. 22, p. 3439, 2023.

[7] M. P. Behera, A. Sarangi, D. Mishra, and S. K. Sarangi, "A hybrid machine learning algorithm for heart and liver disease prediction using modified particle swarm optimization with support vector machine," *Procedia Comput. Sci.*, vol. 218, pp. 818–827, 2023.

[8] H. Gadhe, N. Kolapkar, R. Kakad, and M. N. Sankpal, *Diagnosis of Liver Disease Using Machine Learning*. .

[9] R. Amin, R. Yasmin, S. Ruhi, M. H. Rahman, and M. S. Reza, "Prediction of chronic liver disease patients using integrated projection based statistical feature extraction with machine learning algorithms," *Inform. Med. Unlocked*, vol. 36, no. 101155, p. 101155, 2023.

[10] Z. Jabbar, A. Sattar, A. A. Qusai Al-Neami, and S. M. Khawwam, "Liver fibrosis processing, multiclassification, and diagnosis based on hybrid

machine learning approaches," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 29, no. 3, pp. 1614–1622, 2023.

[11] R. Valeriu, A. Mogos, A.-M. Nechita, G. Adam, A.-S. Adam, and M. Melinte-Popescu, "Machine learning approaches for the prediction of hepatitis B and C seropositivity," *International journal of environmental research and public health*, vol. 20, no. 3, 2023.

[12] Z. Zhao, Y. Tian, Z. Yuan, P. Zhao, F. Xia, and S. Yu, "A machine learning method for improving liver cancer staging," *J. Biomed. Inform.*, vol. 137, no. 104266, p. 104266, 2023.

[13] U. K. Lilhore *et al.*, "Hybrid model for precise hepatitis-C classification using improved random forest and SVM method," *Sci. Rep.*, vol. 13, no. 1, 2023.

[14] L. Sanchez *et al.*, "A machine learning-based classification of adult-onset diabetes identifies patients at risk of liver-related complications," *JHEP Reports*, vol. 5, no. 8, 2023.

[15] T. Jafari, S. Nasiri, M. Sayadi, H. Emami, and S. Mohammadpour, "A Neonatal jaundice prediction system based on the support vector machine algorithm," *Journal of Health Administration*, vol. 25, no. 4, 2023.

[16] S. Listopad *et al.*, "Differentiating between liver diseases by applying multiclass machine learning approaches to transcriptomics of liver tissue or blood-based samples," *JHEP Rep.*, vol. 4, no. 10, p. 100560, 2022.

[17] G. L.-H. Wong *et al.*, "Novel machine learning models outperform risk scores in predicting hepatocellular carcinoma in patients with chronic viral hepatitis," *JHEP Rep.*, vol. 4, no. 3, p. 100441, 2022.

[18] P. R. Kshirsagar, D. H. Reddy, M. Dhingra, D. Dhabliya, and A. Gupta, "Detection of liver disease using machine learning approach," in *2022 5th International Conference on Contemporary Computing and Informatics (IC3I)*, 2022.

[19] S. Tzuu-Hseng, -Jung Huan, and P.-H. Chiu, "Hepatitis C virus detection model by using random forest, logistic-regression and ABC algorithm," *IEEE Access*, vol. 10, pp. 91045–91058, 2022.

[20] O. Oladimeji, A. Oyebisi, and O. Oladimeji, “Machine learning models for diagnostic classification of hepatitis C tests,” *Frontiers in Health Informatics*, vol. 10, no. 1, 2021.

[21] M. G. Ragu, “Anticipation of living status of Hepatitis B patient by using Machine Learning,” *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 9, no. 11, pp. 924–930, 2021.

[22] M. Alghobiri, H. U. Khan, and A. Mahmood, “An empirical comparative analysis using machine learning techniques for liver disease prediction,” *Int. J. Healthc. Inf. Syst. Inform.*, vol. 16, no. 4, pp. 1–12, 2021.

[23] A. A. Kashif, B. Bakhtawar, A. Akhtar, S. Akhtar, N. Aziz, and M. S. Javeid, “Treatment response prediction in hepatitis C patients using machine learning techniques,” *Int. J. TIM*, vol. 1, no. 2, pp. 79–89, 2021.

[24] B. S. Surya, N. K. Singh, and S. S. Rekha, “Implementation of Liver Disease Prediction Using Machine Learning,” *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, pp. 606–616, 2021.

[25] N. K. Trivedi, S. Simaiya, U. K. Lilhore, and S. K. Sharma, “COVID-19 pandemic: Role of machine learning & deep learning methods in diagnosis,” *Int. J. Curr. Res. Rev.*, pp. 150–155, 2021.

[26] L. Syafaâ, Z. Zulfatman, I. Pakaya, and M. Lestandy, “Comparison of machine learning classification methods in hepatitis C virus,” *Jurnal Online Informatika*, vol. 6, no. 1, pp. 73–78, 2021.

[27] M. B. Butt *et al.*, “Diagnosing the Stage of Hepatitis C using machine learning,” *J. Healthc. Eng.*, vol. 2021, p. 8062410, 2021.

[28] J. Singh, S. Bagga, and R. Kaur, “Software-based prediction of liver disease with feature selection and classification techniques,” *Procedia Comput. Sci.*, vol. 167, pp. 1970–1980, 2020.

[29] A. Farzane, M. Akbarzadeh, R. Ferdousi, M. Rashidi, and R. Safdari, “Potential biomarker detection for liver cancer stem cell by machine learning approach,” *J. Contemp. Med. Sci.*, vol. 6, no. 6, pp. 306–312, 2020.

[30] C. Fang, P. Zhang, and X. Qi, “Digital and intelligent liver surgery in the new era: Prospects and dilemmas,” *EBioMedicine*, vol. 41, pp. 693–701, 2019.

[31] M. Nilashi, H. Ahmadi, L. Shahmoradi, O. Ibrahim, and E. Akbari, “A predictive method for hepatitis disease diagnosis using ensembles of neuro-fuzzy technique,” *J. Infect. Public Health*, vol. 12, no. 1, pp. 13–20, 2019.

[32] A. El-Salam *et al.*, “Performance of machine learning approaches on prediction of esophageal varices for Egyptian chronic hepatitis C patients,” *Informatics in Medicine Unlocked*, vol. 17, 2019.

[33] E. L. Godfrey *et al.*, “The decreasing predictive power of MELD in an era of changing etiology of liver disease,” *Am. J. Transplant*, vol. 19, no. 12, pp. 3299–3307, 2019.

[34] A. M. Remita and A. B. Diallo, “Statistical linear models in virus genomic alignment-free classification: Application to hepatitis C viruses,” in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2019.

[35] D. Omran *et al.*, “Towards hepatitis C virus elimination: Egyptian experience, achievements and limitations,” *World J. Gastroenterol.*, vol. 24, no. 38, pp. 4330–4340, 2018.

[36] M. Stepanova, I. Younossi, A. Racila, and Z. M. Younossi, “Prediction of health utility scores in patients with chronic hepatitis C using the chronic liver disease questionnaire-hepatitis C version (CLDQ-HCV),” *Value Health*, vol. 21, no. 5, pp. 612–621, 2018.

[37] J. Cai, T. Chen, and X. Qiu, “Fibrosis and inflammatory activity analysis of chronic hepatitis C based on extreme learning machine,” in *2018 9th International Conference on Information Technology in Medicine and Education (ITME)*, 2018.

[38] R. Wei *et al.*, “Clinical prediction of HBV and HCV related hepatic fibrosis using machine learning,” *EBioMedicine*, vol. 35, pp. 124–132, 2018.

[39] S. Hashem *et al.*, “Comparison of machine learning approaches for prediction of advanced liver fibrosis in chronic hepatitis C patients,” *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 15, no. 3, pp. 861–868, 2018.

[40] N. Chalasani *et al.*, "The diagnosis and management of nonalcoholic fatty liver disease: Practice guidance from the American Association for the Study of Liver Diseases," *Hepatology*, vol. 67, no. 1, pp. 328–357, 2018.

[41] G. M. Emmanuel, T. S. Coshatt, S. Winokur, and S. L. Harada, "Alchemy: A Web 2.0 real-time quality assurance platform for human immunodeficiency virus, hepatitis C virus, and BK virus quantitation assays," *Journal of Pathology Informatics*, vol. 8, no. 1, 2017.

[42] R. Loomba *et al.*, "Gut microbiome-based metagenomic signature for non-invasive detection of advanced fibrosis in human nonalcoholic fatty liver disease," *Cell Metab.*, vol. 25, no. 5, pp. 1054-1062.e5, 2017.

[43] Y. Hayashi and K. Fukunaga, "Accuracy of rule extraction using a recursive-rule extraction algorithm with continuous attributes combined with a sampling selection technique for the diagnosis of liver disease," *Inform. Med. Unlocked*, vol. 5, pp. 26–38, 2016.

[44] T. R. Baitharu and S. K. Pani, "Analysis of data mining techniques for healthcare decision support system using liver disorder dataset," *Procedia Comput. Sci.*, vol. 85, pp. 862–870, 2016.

[45] K. T. Suk and D. J. Kim, "Staging of liver fibrosis or cirrhosis: The role of hepatic venous pressure gradient measurement," *World J. Hepatol.*, vol. 7, no. 3, pp. 607–615, 2015.

[46] N. Méndez-Sánchez *et al.*, "Latin American association for the study of the liver recommendations on treatment of hepatitis C," *Ann. Hepatol.*, vol. 13, pp. S4–S66, 2014.

[47] A. H. KayvanJoo, M. Ebrahimi, and G. Haqshenas, "Prediction of hepatitis C virus interferon/ribavirin therapy outcome based on viral nucleotide attributes using machine learning algorithms," *BMC Res. Notes*, vol. 7, no. 1, p. 565, 2014.

[48] M. ElHefnawi *et al.*, "Accurate prediction of response to interferon-based therapy in Egyptian patients with chronic hepatitis C using machine-learning approaches," in *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2012.

[49] W. Shi, I. T. Freitas, C. Zhu, W. Zheng, W. W. Hall, and D. G. Higgins, “Recombination in hepatitis C virus: Identification of four novel naturally occurring inter-subtype recombinants,” *PLoS One*, vol. 7, no. 7, p. e41997, 2012.

[50] E. F. Duffell, M. J. W. Van De Laar, and A. J. Amato-Gauci, “Enhanced surveillance of hepatitis C in the EU, 2006–2012,” *Journal of Viral Hepatitis*, vol. 22, no. 7, pp. 590–595, 2006.

plagarism rep.docx

ORIGINALITY REPORT

13%	8%	8%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Amir Shachar. "Introduction to Algogens", Open Science Framework, 2024 Publication	1%
2	www.mdpi.com Internet Source	1%
3	www2.mdpi.com Internet Source	1%
4	link.springer.com Internet Source	<1%
5	www.frontiersin.org Internet Source	<1%
6	www.ijesr.org Internet Source	<1%
7	Junjie Wu, RenFu Yang, Peng Zhao, LuXia Yang. "Computer-aided mobility solutions: Machine learning innovations to secure smart urban transportation", Sustainable Cities and Society, 2024 Publication	<1%