Modelling Treatment Outcomes - Liver

Muhammad Ehtisham Hassan and Rayyan Shabbir

CureMD

Abstract. Liver disease is a significant global health challenge, responsible for millions of deaths annually. Key causes include excessive alcohol consumption, hepatitis infections, and other metabolic conditions like obesity. Predicting liver disease is crucial for early diagnosis and effective treatment. With the advent of machine learning (ML) models, significant progress has been made in automating diagnosis and predicting patient outcomes. This paper evaluates several machine learning models, such as Random Forest (RF), XGBoost, Support Vector Machine (SVM), Logistic Regression (LR), and AdaBoost, for their effectiveness in predicting liver disease progression using a dataset that includes key health indicators, such as ALT, AST, albumin levels, and other vital parameters.

The study uses a dataset of over 500,000 patients, where data filtering, feature extraction, and cleaning procedures were applied to remove inconsistencies like missing lab results, null vital values, and outliers in diagnostic measures. These preprocessing steps included the imputation of missing demographic values, consistency checks for lab measurements, and multi-hot encoding for diagnoses. Key health indicators, including ALT and AST, were selected as target variables to assess liver functionality. The dataset was also structured to include the patient's demographic information, vitals, diagnosis, and medications administered during the treatment period.

After data preprocessing, several models were trained to predict liver disease outcomes, including regression models and ensemble methods like RF and Gradient Boosting. The performance of these models was evaluated using metrics such as accuracy, precision, recall, F1 score, and AUC. Random Forest emerged as the most effective model with high accuracy, demonstrating its robustness in handling complex health datasets and its ability to reduce overfitting by building multiple decision trees.

This study not only presents insights into the modeling process but also highlights the practical implications of utilizing machine learning in healthcare settings. The models' predictive accuracy can aid clinicians in diagnosing liver disease earlier, allowing for more personalized treatment plans. The integration of ML models into routine diagnostic processes could lead to better patient outcomes and reduced healthcare costs by improving early detection rates.

Keywords: Machine Learning · Liver Health Prediction.

1 Introduction

Liver disease poses a significant public health challenge worldwide. Chronic liver disease (CLD) is often caused by factors such as excessive alcohol consumption, viral infections (like hepatitis), obesity, and metabolic syndrome. It is frequently asymptomatic in its early stages, making early diagnosis and intervention crucial for effective treatment. Globally, liver disease accounts for approximately 2 million deaths annually, and in the United States alone, it ranks among the leading causes of death. Given the complexity and variability of liver conditions, identifying the risk factors that influence liver disease progression is key to improving outcomes.

Machine learning (ML) has emerged as a promising tool in healthcare for automating disease diagnosis and predicting patient outcomes. By leveraging large datasets and advanced algorithms, ML models can uncover patterns that might not be evident to human clinicians. These models can assist in diagnosing diseases at earlier stages, improving treatment options, and reducing the overall burden of liver disease.

This paper presents a comprehensive approach to predicting liver disease outcomes using several ML models, including Random Forest, XGBoost, Support Vector Machine, Logistic Regression, and AdaBoost. By analyzing a dataset of over 500,000 patients, we aim to identify key health factors and develop predictive models that can assist in diagnosing liver disease early. The study also investigates the challenges associated with data preprocessing, feature extraction, and model training in a healthcare context. Through these efforts, we aim to contribute to the growing body of knowledge on how ML can be applied to improve liver disease diagnosis and treatment.

2 Literature Review

Liver disease is a significant global health issue, contributing to millions of deaths annually due to conditions such as cirrhosis, hepatitis, and liver cancer. The early detection and management of liver disease are crucial for improving patient outcomes and reducing healthcare costs. In recent years, machine learning (ML) techniques have emerged as powerful tools in medical diagnosis and prediction, offering enhanced accuracy and efficiency over traditional methods. The application of ML models to liver disease prediction has garnered increasing attention, with a variety of algorithms being tested and refined to improve diagnostic capabilities. This literature review examines the current landscape of machine learning approaches for liver disease prediction, focusing on commonly used models, their effectiveness, and the limitations observed in practical applications. Additionally, it highlights the gaps in existing research and discusses potential improvements to address these challenges.

2.1 Approach

The approach for this literature review involves the analysis of 10 research papers focused on liver disease prediction using machine learning (ML) models. The aim is to identify the strengths and limitations of existing models and to explore how machine learning can improve liver disease diagnosis and prediction, particularly in managing the challenges associated with chronic diseases like liver disease.

2.2 Database Search

The studies included in this review were identified through searches in academic databases such as PubMed, IEEE Xplore, and ResearchGate. Search terms like ”liver disease prediction,” ”machine learning for liver disease,” ”chronic liver disease ML models,” and ”ensemble learning liver disease” were used to find relevant papers published in the last five years. A total of 10 papers were selected, providing a comprehensive view of different machine learning techniques for liver disease prediction.

2.3 Inclusion Criteria

The selected studies were required to meet the following criteria:

- Focus on liver disease prediction using machine learning techniques.
- Use of public or clinical datasets, such as the Indian Liver Patient Dataset (ILPD) or clinical data from hospitals.
- Evaluation of model performance using metrics such as accuracy, precision, recall, F1 score, and area under the ROC curve (AUC).
- Papers published between 2019 and 2024 to ensure up-to-date findings.

2.4 Exclusion Criteria

Studies were excluded if they:

- Focused on liver disease diagnosis without applying machine learning models.
- Did not report performance metrics or validation results.
- Used proprietary datasets without publicly available data or details.

2.5 Data Extraction

Data were extracted from the selected studies, focusing on:

- Dataset description (e.g., number of patients, features, and target variables).
- Machine learning models used (e.g., random forest, decision tree, logistic regression).
- Performance evaluation metrics (e.g., accuracy, precision, recall, F1 score, AUC).
- Challenges encountered in liver disease prediction.

2.6 Limitations of Other Machine Learning Models in Liver Health Prediction

Many machine learning models used in liver health prediction face several limitations that affect their performance and utility in clinical settings. One major challenge is the limited feature scope; many models focus on a small set of features, such as basic demographic data or liver function tests, without considering other relevant factors like genetics, lifestyle, or environmental influences. Additionally, most models treat patient data as static, failing to account for the temporal progression of liver disease. This limitation can lead to inaccurate predictions, especially in chronic conditions where liver health deteriorates over time. Handling missing data is another common issue, as many models struggle when data is incomplete, and imputation techniques can introduce bias. Overfitting is also prevalent, particularly in models trained on small or imbalanced datasets, leading to high accuracy during training but poor generalization to new data. Finally, while complex models like neural networks and ensemble methods offer high accuracy, they often lack interpretability, making it difficult for healthcare professionals to trust and implement these predictions in clinical practice. Addressing these limitations requires a more holistic approach to model design, incorporating broader feature sets, handling temporal data, improving missing data techniques, and increasing model interpretability.

2.7 Limited Feature Scope

One major limitation observed in the papers is the use of a limited feature scope. Many models rely on a small set of features, such as basic liver function test results or demographic information, which can limit the predictive power of the model. For example, studies like those by Ganie et al. (2024) focused on a few biochemical markers, which may overlook other important factors such as lifestyle and comorbidities. Solution:Expanding the feature set to include a broader range of data, such as genetic markers, environmental exposures, and patient history, could improve the accuracy of liver disease prediction models.

2.8 Static Models and Lack of Temporal Consideration

Most models are static and do not account for temporal changes in patient data, such as trends in liver enzyme levels over time. This static approach fails to capture the progression of liver diseases or the response to treatments. Solution:Implementing dynamic, time-series models like recurrent neural networks (RNNs) or long short-term memory (LSTM) networks could provide better insights into how liver diseases progress over time.

2.9 Insufficient Handling of Missing Data

Missing data is a common issue in healthcare datasets, particularly when dealing with liver disease. Many studies applied imputation techniques but did not account for the uncertainty introduced by missing values. Solution:More robust approaches, such as multiple imputations or advanced techniques like deep learning models that can handle missing data directly, should be applied. The use of Bayesian models can also help in quantifying the uncertainty associated with missing data.

2.10 Overfitting on Small or Imbalanced Datasets

A number of studies reported overfitting problems, particularly in small or imbalanced datasets where the number of positive liver disease cases is significantly higher than negative ones. Models such as decision trees and neural networks tend to memorize the training data rather than generalize well to new, unseen data. Solution:Addressing this requires the use of techniques like cross-validation, regularization, and the application of ensemble methods like bagging and boosting to reduce overfitting.

2.11 Lack of Interpretability in Complex Models

Models like deep learning and ensemble methods, while highly accurate, often suffer from a lack of interpretability. This can make it difficult for clinicians to trust the model's predictions and integrate them into clinical practice. Solution:Incorporating interpretable machine learning techniques, such as SHAP values or LIME, can help explain model decisions, thereby increasing clinician trust and understanding.

2.12 Generalization Issues Across Different Populations

Several studies identified that models trained on specific population datasets often fail to generalize well to other populations due to variations in demographic factors, genetic background, and healthcare practices. Solution:Creating more diverse and inclusive datasets that represent different populations could improve model generalization. Transfer learning techniques can also be applied to adapt models trained on one population to new, diverse populations.

2.13 Ignoring Medication Side Effects and Interactions

Few models consider the effects of medications, particularly their side effects and interactions with liver functions. This omission can lead to incorrect predictions, especially for patients with polypharmacy. Solution:Future models should incorporate medication data, including dosages and known side effects, into the feature set to improve prediction accuracy.

2.14 Inadequate Performance Evaluation and Validation

Some studies lack robust performance evaluation methods, using only basic accuracy metrics without considering other important measures like precision, recall, or AUC. This makes it difficult to assess model performance effectively. Solution:Comprehensive evaluation frameworks using multiple performance metrics and external validation sets should be adopted to ensure models are both accurate and clinically relevant.

2.15 Scalability and Deployment Challenges

Finally, several studies identified difficulties in scaling and deploying machine learning models in real-world clinical settings due to the computational complexity and the need for specialized hardware. Solution:Leveraging cloud-based services and optimizing model complexity for deployment on standard hospital hardware could address scalability issues.

2.16 Synthesis of Findings

The synthesis of findings reveals that while machine learning models offer promising advancements in liver disease prediction, there are notable challenges that need to be addressed for broader clinical applicability. Many models achieve high accuracy, especially ensemble methods like gradient boosting and random forests; however, limitations arise due to factors like limited feature scope, overfitting, and lack of interpretability. Static models often miss temporal trends crucial for chronic disease progression, and handling missing data remains a challenge that can skew predictions. Furthermore, models trained on specific populations sometimes struggle to generalize across diverse demographic groups. Addressing these limitations—through broader feature inclusion, dynamic modeling, robust data handling, and interpretable algorithms—could enhance the effectiveness and reliability of machine learning applications in liver health prediction.

- Regression Models: Linear regression, logistic regression
- Classification Models: Decision trees, random forests, support vector machines (SVM), k-nearest neighbors (KNN), Naive Bayes
- Time-Series Analysis: ARIMA, LSTM (Long Short-Term Memory networks)
- Ensemble Methods: Gradient boosting, bagging, stacking, AdaBoost, XGBoost
- Neural Networks: Artificial neural networks (ANNs), multilayer perceptrons (MLP)

2.17 Role of Medications

Medications, including their side effects and interactions with liver function, are often overlooked in liver disease prediction models. Integrating detailed medication data can lead to more accurate predictions, especially in patients undergoing multiple treatments.

2.18 Model Performances in Literature Review

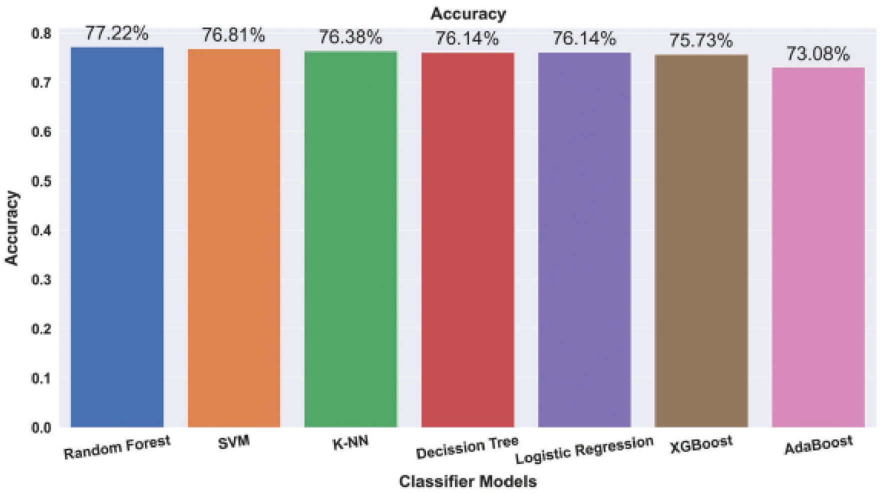
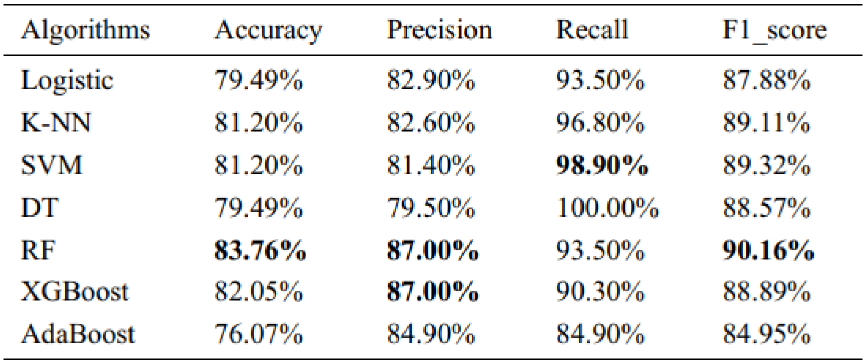


Fig. 1: Average Accuracy of applied algorithms in various train sizes

Table 2: Statistical results (80% train size) of this study



Algorithms	Accuracy	Precision	Recall	F1_score
Logistic	79.49%	82.90%	93.50%	87.88%
K-NN	81.20%	82.60%	96.80%	89.11%
SVM	81.20%	81.40%	**98.90%**	89.32%
DT	79.49%	79.50%	100.00%	88.57%
RF	**83.76%**	**87.00%**	93.50%	**90.16%**
XGBoost	82.05%	**87.00%**	90.30%	88.89%
AdaBoost	76.07%	84.90%	84.90%	84.95%

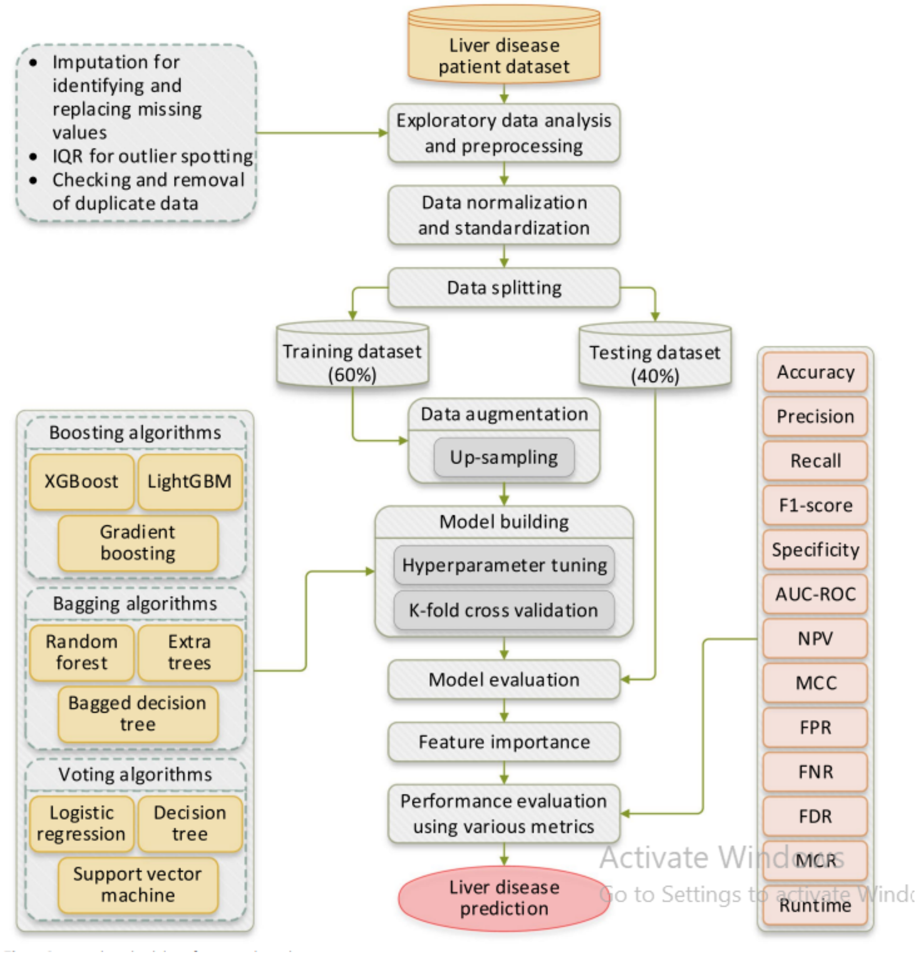


Fig. 2: Proposed Methodology in Literature

3 Methodology

This section outlines the process of data extraction, exploratory data analysis (EDA), and modeling for predicting liver disease outcomes using machine learning models. The methodology is divided into several subsections, each focusing on the key stages of data handling and model development. Data from various sources, including lab results, demographics, vitals, diagnoses, and medications, was filtered, cleaned, and analyzed to build effective models. Below, we describe each step of the methodology in detail.

3.1 Lab Pair Selection

By plotting the differences between all consecutive labs for all patients, the optimal timeframe was found to be 3 months. The selected labs for all patients are between 80-100 days apart.

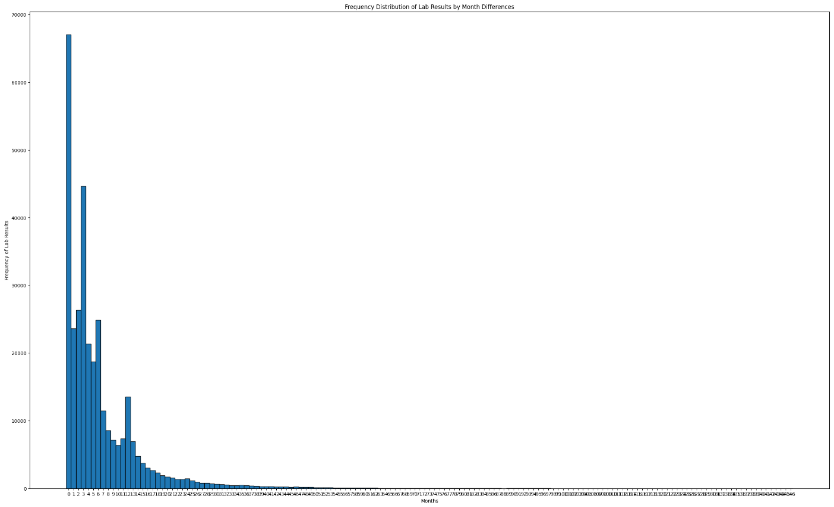
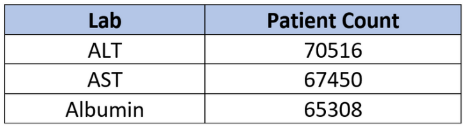


Fig. 3: Frequency distribution of lab results by month differences

3.2 Extraction and EDA of Lab Results

The dataset used in this study included lab results for key biomarkers related to liver health, such as ALT (Alanine Aminotransferase), AST (Aspartate Aminotransferase), and albumin levels. These biomarkers are widely recognized as indicators of liver functionality. Elevated ALT and AST levels are linked to liver damage, while albumin levels reflect the liver's ability to synthesize proteins.



Lab	Patient Count
ALT	70516
AST	67450
Albumin	65308

Fig. 4: Patient counts for each lab

Step 1: Data Filtering and Cleaning The raw dataset included inconsistencies in lab values, such as missing data, incorrectly recorded units, and outliers.

For example, units for ALT and AST were reported as U/L but sometimes included erroneous entries, such as '61' instead of the correct unit. Patients with lab values listed as 'NaN' or nonsensical entries like 'Pending' were excluded from the analysis. Additionally, outliers were removed to improve model performance. For ALT and AST, only patients with values between 1-120 IU/L were included, as levels higher than this typically indicate acute liver disease and were deemed outliers.

Step 2: Binning of Lab Values To improve model performance and address class imbalance, lab values for ALT and AST were binned into multiple categories based on clinical guidelines. Initially, five bins were created for ALT, but this resulted in significant class imbalances. To mitigate this, the values were re-binned into 10 categories. However, the model performance remained suboptimal, leading to the decision to switch to regression-based modeling for more nuanced predictions.

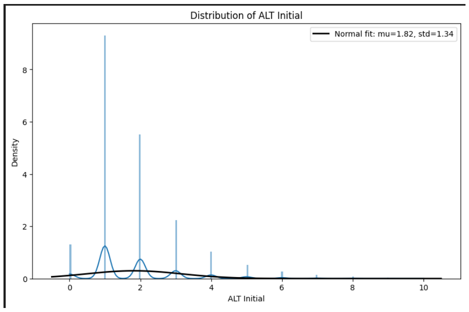


Fig. 5: Distribution of ALT

Step 3: Exploratory Data Analysis (EDA) The EDA process began with plotting the distribution of lab results to understand the spread and identify anomalies. Correlation matrices were generated to observe the relationships between ALT, AST, and albumin levels. This analysis helped refine the feature set by focusing on the most relevant indicators of liver disease. For instance, a strong positive correlation between ALT and AST indicated their mutual relevance in liver disease progression, leading to their inclusion as key features.

3.3 Extraction and EDA of Demographics

Demographic data plays a critical role in predicting liver disease outcomes as factors such as age, gender, and ethnicity can influence disease progression. For instance, older age is a known risk factor for chronic liver disease, and certain ethnic groups have higher rates of conditions like hepatitis.

Step 1: Gender and Age Mapping The demographic dataset included variables like gender and age. Gender was encoded as a binary variable, with 'F' mapped to 0 and 'M' mapped to 1. Age was computed from the patients' birthdates and later binned into age groups for more precise analysis, reflecting typical medical risk categories.

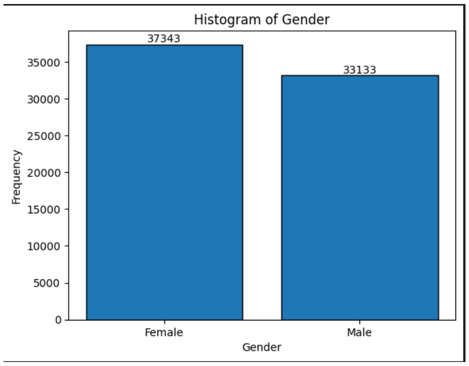


Fig. 6: Distribution of Gender

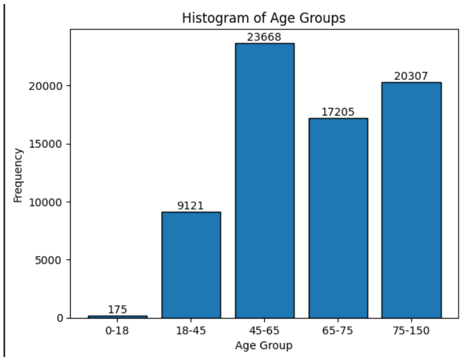


Fig. 7: Distribution of Age

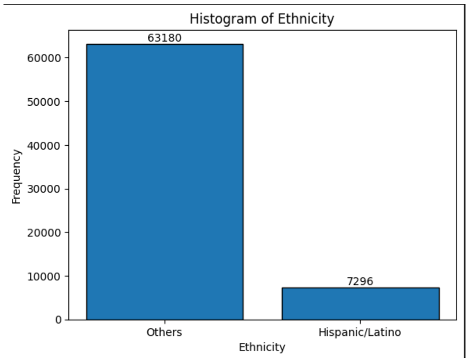


Fig. 8: Distribution of Ethnicity

Step 2: Handling Missing Demographic Data A significant challenge was dealing with missing demographic values, particularly ethnicity. Approximately 16,800 entries in the ethnicity column were missing. To handle this, the mode of the column (the most frequently occurring value) was used for imputation, a common practice when the missingness is assumed to be random.

Step 3: Correlation Analysis and EDA The correlation between demographic variables and liver disease outcomes was explored to understand the role of these factors in disease progression. For example, older patients showed a higher correlation with elevated ALT levels, which is consistent with the literature on age-related liver degeneration. The distribution of patients across gender and ethnic groups was also visualized to ensure that the dataset was representative of a broad population, minimizing bias in the predictive models.

3.4 Extraction and EDA of Vitals

Vitals such as body mass index (BMI), height, weight, and mean arterial pressure (MAP) are often indicative of liver health. These values were extracted and analyzed to identify trends and ensure consistency.

Step 1: Conversion of Units Vitals were reported in different units across the dataset. For example, height was sometimes recorded in centimeters and sometimes in inches. These discrepancies were resolved by converting all values to standard units (SI units) before further analysis. Additionally, BMI was calculated from height and weight for each patient to provide a more comprehensive measure of overall health.

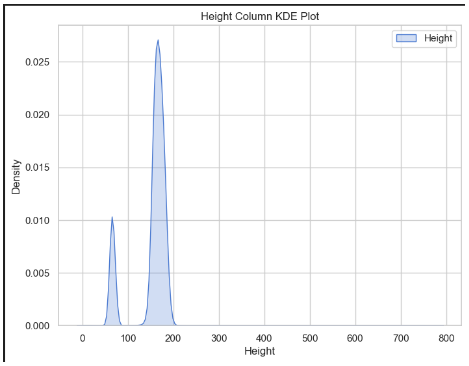


Fig. 9: Distribution of Height

Step 2: Handling Missing and Inconsistent Data Many patients had missing or inconsistent vitals. For example, null values for weight or height were common. These missing values were imputed using the median of the available data, as median imputation is robust to outliers. Extreme outliers, such as a BMI below 10 or above 50, were removed, as these values are medically implausible and could distort model predictions.

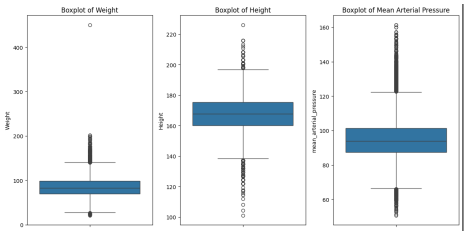


Fig. 10: Boxplots

Step 3: Exploratory Data Analysis Once the vital data was cleaned, we conducted an EDA to assess the distribution of key vitals across different patient groups. For instance, patients with liver disease generally had higher BMIs, consistent with existing literature that links obesity with liver conditions such as fatty liver disease. MAP, a key indicator of cardiovascular health, was also analyzed to understand its relationship with liver function, as impaired liver function is often associated with increased cardiovascular risks.

3.5 Extraction and EDA of Diagnosis

Diagnoses were extracted from the dataset using ICD codes to provide insight into the patients' medical history. These codes represent various conditions, some of which are directly related to liver health, while others may affect the liver indirectly.

Step 1: Diagnosis related to Liver In this analysis, we focused exclusively on health conditions with a rationale score of 2 or higher (section 3.5), as these conditions exhibit at least an indirect impact on liver health. A rationale score of 2 suggests potential relevance through metabolic or systemic interactions, while higher scores indicate a stronger, more direct association with liver function or disease progression. By filtering for conditions with these scores, we aim to highlight those health issues that could reasonably contribute to or exacerbate liver-related health outcomes, ensuring the analysis remains focused on meaningful associations.

Step 2: Multi-hot Encoding of ICD Codes Initially, diagnoses were encoded using multi-hot encoding, which converts each diagnosis into a binary variable. This process resulted in over 1,370 unique columns, one for each diagnosis code. To simplify the dataset, ICD codes that appeared in fewer than 2,000 patients were removed.

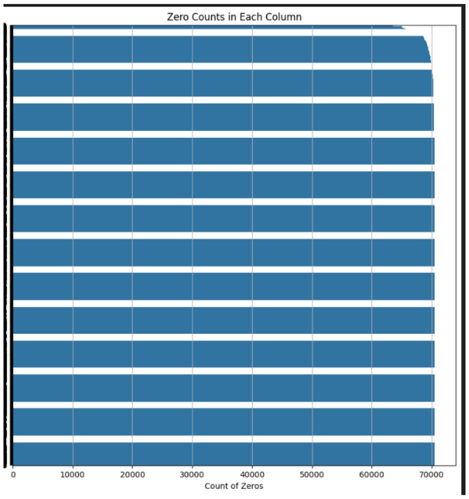


Fig. 11: Zero Counts of Columns

Step 3: Grouping of ICD Codes To further reduce complexity, ICD codes were grouped into broader categories. This grouping was based on clinical expertise and the Elixhauser Comorbidity Index, which is widely used in healthcare to group diagnoses into comorbid conditions. For instance, codes related to hypertension, diabetes, and cardiovascular disease were grouped together. The

Table 1: Health Conditions with Potential Impact on Liver Health.

ICD Code	Description	Liver Health Impact (1-5)	Rationale
C50	Malignant neoplasm of breast	1	No direct relevance to liver health.
D50	Iron deficiency anemia	1	Generally unrelated to liver conditions.
D64	Anemia, unspecified	1	Not directly linked to liver health.
E03	Hypothyroidism, unspecified	1	No significant connection to liver function.
E11	Type 2 diabetes mellitus	2	Can contribute to non-alcoholic fatty liver disease (NAFLD), but not directly indicative of liver disease.
E29	Testicular hypofunction	1	No relation to liver health.
E53	Vitamin deficiency, unspecified	2	Certain deficiencies can affect liver health indirectly.
E55	Vitamin D deficiency	2	May have indirect effects on liver health but not directly related.
E66	Obesity	3	Strongly associated with NAFLD and other liver conditions.
E78	Disorders of lipoprotein metabolism	3	Can indicate metabolic syndrome, which is linked to liver disease.
F32	Major depressive disorder	1	No direct impact on liver health.
F41	Anxiety disorders	1	Not related to liver function.
G47	Sleep disorders	1	No significant connection to liver health.
I10	Essential hypertension	2	Hypertension can be associated with metabolic syndrome affecting the liver indirectly.
I25	Chronic ischemic heart disease	2	Indirectly related through overall metabolic health but not specific to the liver.
J30	Allergic rhinitis	1	No relation to liver conditions.
K21	Gastro-esophageal reflux disease	3	While primarily gastrointestinal, chronic reflux may relate to lifestyle factors affecting liver health.
M25	Other joint disorders	1	Not relevant to liver conditions.
M54	Dorsalgia (back pain)	1	No direct connection to liver health.
M79	Other soft tissue disorders	1	Not related to the liver.
M81	Osteoporosis without current pathological fracture	1	No direct relevance to the liver.
N18	Chronic kidney disease	3	Kidney and liver diseases often coexist; however, this code is more kidney-focused.
N39	Other disorders of urinary system	1	Not relevant to liver health.
N40	Benign prostatic hyperplasia	1	No direct connection to the liver.
R00	Abnormalities of heart rhythm	2	May reflect overall metabolic issues that could indirectly affect the liver.
R05	Cough	1	No relevance to liver conditions.
R06	Abnormalities of breathing	1	Not related to the liver's functionality or health.
R07	Pain in throat and chest	1	No direct impact on liver health.
R10	Abdominal and pelvic pain	3	Potentially indicative of underlying abdominal issues that could involve the liver.
R53	Malaise and fatigue	2	General symptoms that could relate indirectly to various health issues, including the liver.
R63	Symptoms and signs concerning food and fluid intake	2	Could indicate nutritional issues affecting overall health, including the liver indirectly.
R73	Elevated blood glucose level	3	Strongly associated with metabolic syndrome and potential NAFLD risk factors.
R79	Abnormal findings on examination of blood chemistry	2	May indicate underlying issues that could involve the liver but is not specific enough for direct correlation.
R94	Abnormal findings on diagnostic imaging	2	Could relate indirectly to various organ systems, including the liver, depending on findings.

grouped data was then used to explore how comorbid conditions influence liver health.

Step 4: Correlation Analysis Correlation matrices were used to identify relationships between various diagnoses and liver disease outcomes. Conditions such as diabetes and cardiovascular disease showed a strong positive correlation with elevated ALT and AST levels, reinforcing the well-established connection between metabolic syndrome and liver disease progression.

3.6 Extraction and EDA of Medications

Medications administered during the treatment period were another crucial factor influencing liver outcomes. Some medications, such as statins or antivirals, directly affect liver enzymes, while others may have indirect effects.

Step 1: Filtering Medications by Time Frame The dataset included records of medications taken by patients over different time periods. To focus on the most relevant medications, only those administered during the time frame of the patient's lab results were retained. Medications taken outside of this period were excluded to reduce noise in the analysis.

Step 2: Grouping of Medications Medications were grouped based on their GPI codes, which categorize drugs by their therapeutic use. Medications that directly impact liver function were prioritized, while those unrelated to liver health were excluded from further analysis. For example, drugs affecting lipid metabolism were included due to their known influence on liver enzymes.

Step 3: Analysis of Medication Effects Once the medication data was cleaned and grouped, an EDA was conducted to explore the effects of different medications on liver health. For instance, patients on statins generally had lower ALT and AST levels, indicating that lipid-lowering drugs may have a protective effect on liver function. The relationship between medication use and liver health was further explored using correlation matrices and visualization tools.

4 Modelling

The final step in the methodology involved training and evaluating machine learning models to predict liver disease outcomes. Multiple models were tested, including Random Forest, XGBoost, and Gradient Boosting Regressor, each chosen for their ability to handle large datasets and their robustness in medical applications.

4.1 Model Selection and Hyperparameter Tuning

Random Forest (RF) was selected for its effectiveness in handling both categorical and continuous variables. RF is particularly useful for medical datasets as it reduces overfitting by averaging the results of multiple decision trees. Similarly, XGBoost and Gradient Boosting were chosen for their ability to handle imbalanced datasets and improve prediction accuracy through iterative refinement.

4.2 Challenges

The primary challenges encountered during modeling included class imbalance, missing data, and overfitting. For example, the dataset had an unequal distribution of patients across liver disease stages, which made it difficult for the models to learn from the minority classes. To address this, techniques like oversampling and weighted loss functions were applied to balance the dataset during model training.

4.3 Evaluation Metrics

The models were evaluated using metrics such as accuracy, precision, recall, F1 score, and area under the ROC curve (AUC). Random Forest emerged as the most accurate model, achieving an accuracy of over 80 percent, while XGBoost and Gradient Boosting performed slightly lower but still provided valuable insights. Hyperparameter tuning was performed using grid search, which further improved model performance by identifying the optimal parameters for each algorithm

5 Sequence Diagram

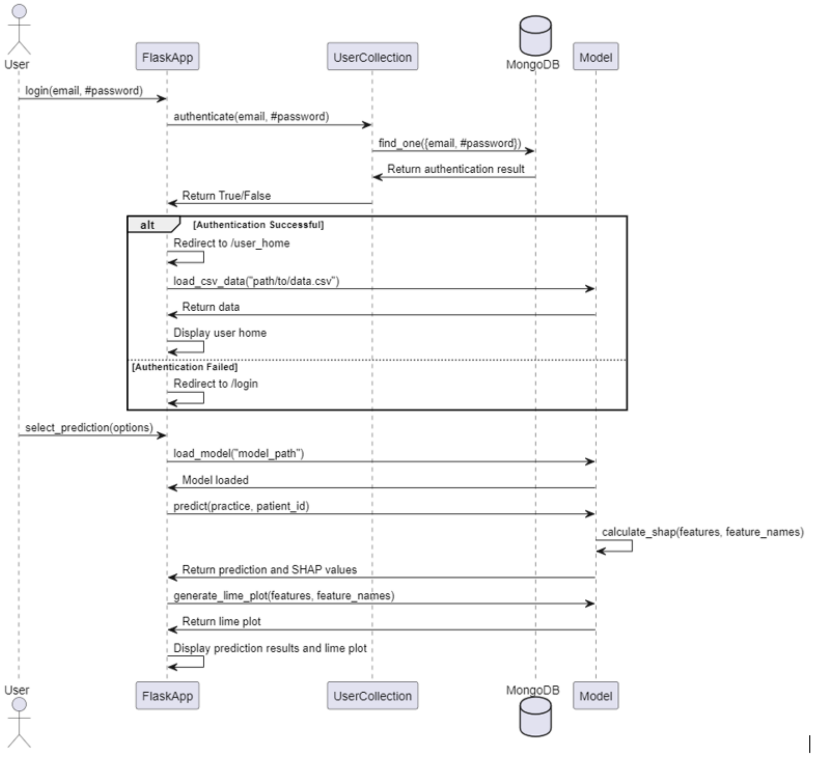

User
FlaskApp
UserCollection
MongoDB
Model
login(email, #password)
authenticate(email, #password)
find_one({email, #password})
Return authentication result
Return True/False
alt
[Authentication Successful]
Redirect to /user_home
load_csv_data("path/to/data.csv")
Return data
Display user home
[Authentication Failed]
Redirect to /login
select_prediction(options)
load_model("model_path")
Model loaded
predict(practice, patient_id)
calculate_shap(features, feature_names)
Return prediction and SHAP values
generate_lime_plot(features, feature_names)
Return lime plot
Display prediction results and lime plot
User
FlaskApp
UserCollection
MongoDB
Model

Fig. 12: Sequence Diagram

6 Architecture Diagram

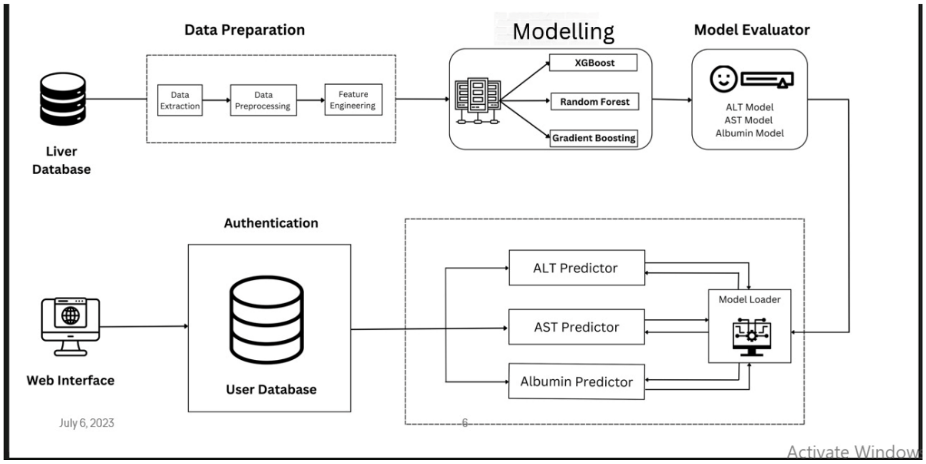
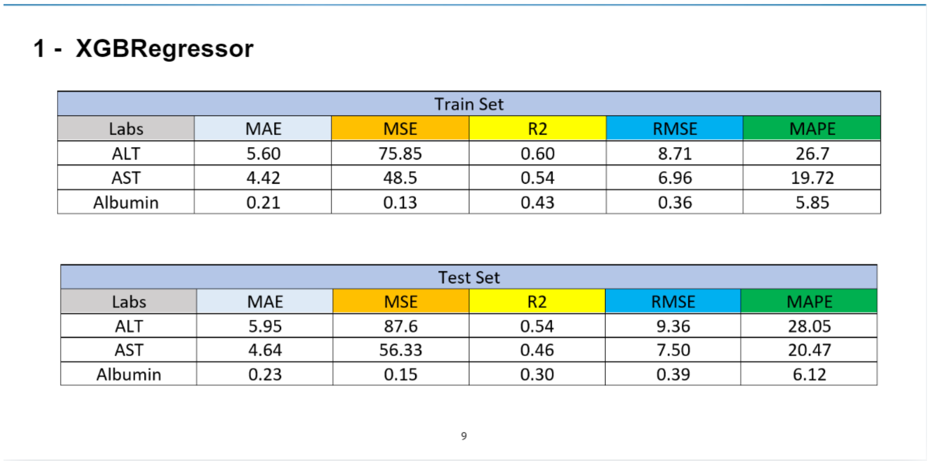


Fig. 13: Architecture Diagram

7 Results



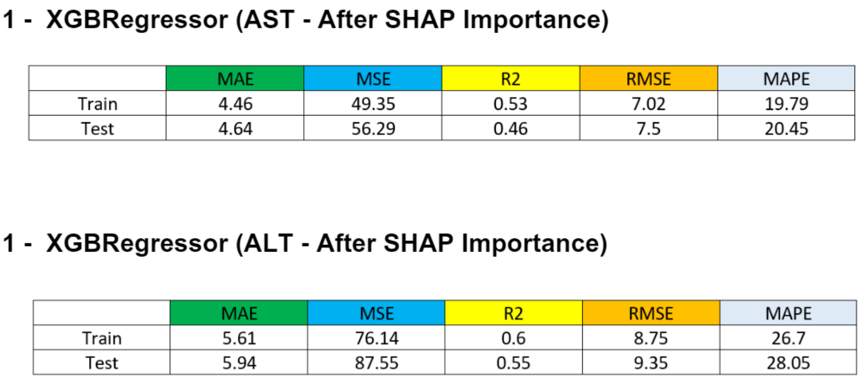
1 - XGBRegressor

Train Set					
Labs	MAE	MSE	R2	RMSE	MAPE
ALT	5.60	75.85	0.60	8.71	26.7
AST	4.42	48.5	0.54	6.96	19.72
Albumin	0.21	0.13	0.43	0.36	5.85

Test Set					
Labs	MAE	MSE	R2	RMSE	MAPE
ALT	5.95	87.6	0.54	9.36	28.05
AST	4.64	56.33	0.46	7.50	20.47
Albumin	0.23	0.15	0.30	0.39	6.12

9

Fig. 14: XGB Regressor



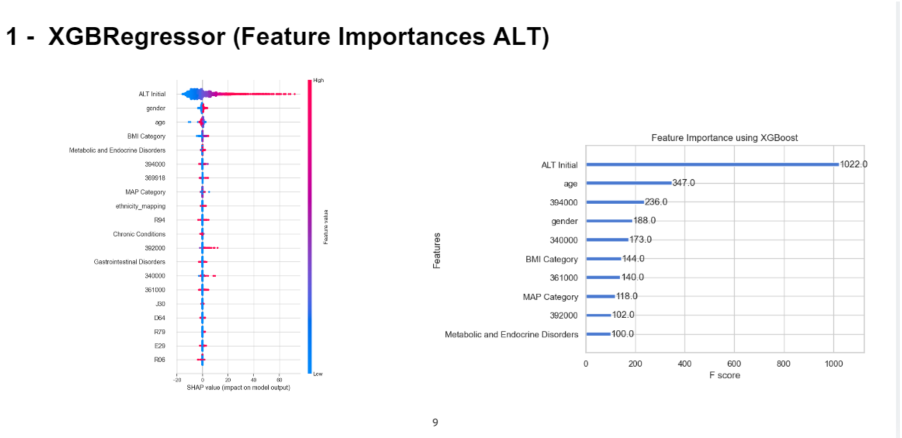
1 - XGBRegressor (AST - After SHAP Importance)

	MAE	MSE	R2	RMSE	MAPE
Train	4.46	49.35	0.53	7.02	19.79
Test	4.64	56.29	0.46	7.5	20.45

1 - XGBRegressor (ALT - After SHAP Importance)

	MAE	MSE	R2	RMSE	MAPE
Train	5.61	76.14	0.6	8.75	26.7
Test	5.94	87.55	0.55	9.35	28.05

Fig. 15: XGB Regressor after removal of noise features using SHAP



1 - XGBRegressor (Feature Importances ALT)

Fig. 16: XGB Regressor Feature Importance for ALT

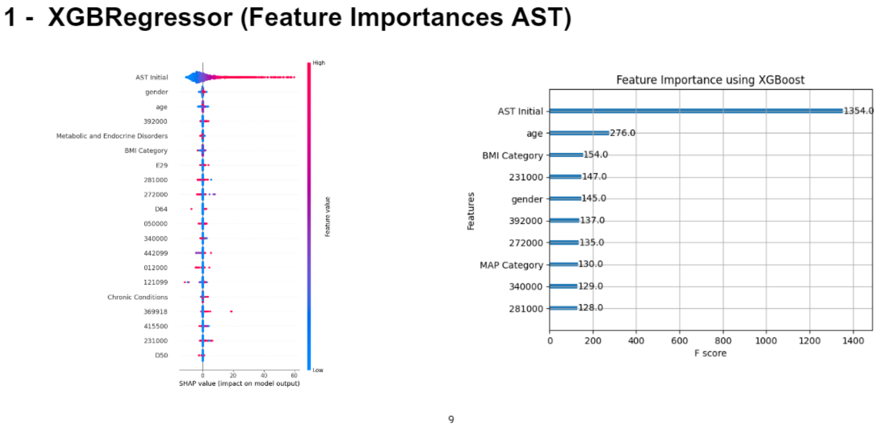


Fig. 17: XGB Regressor Feature Importance for ALT

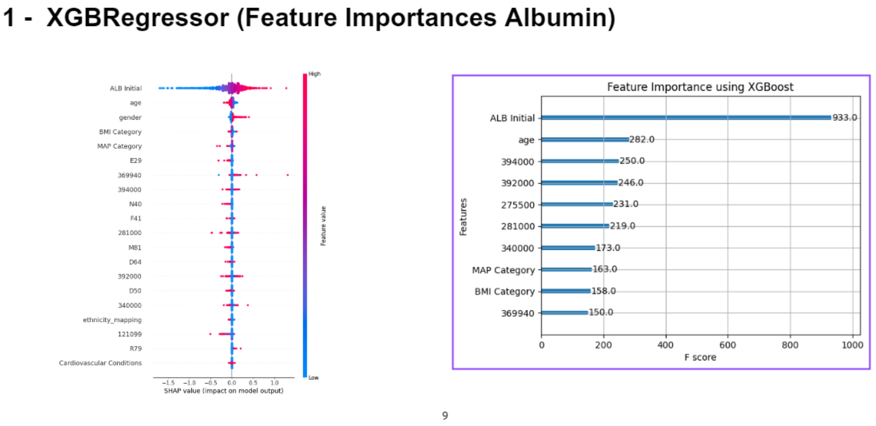
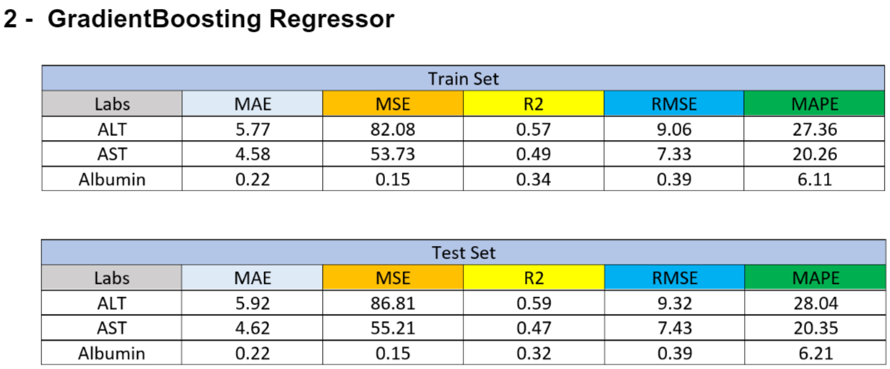


Fig. 18: XGB Regressor Feature Importance for ALT

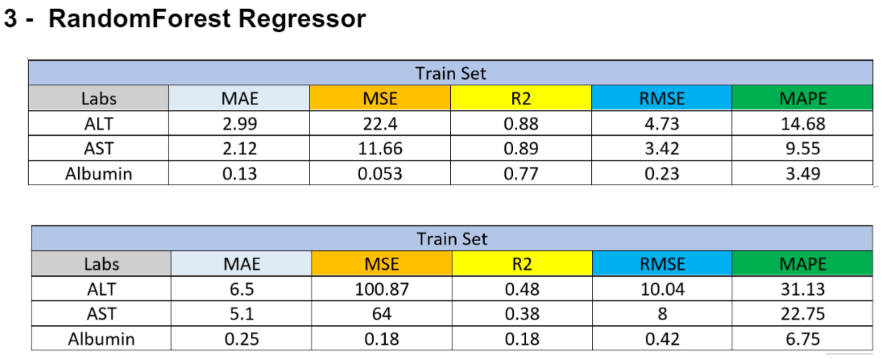


2 - GradientBoosting Regressor

Train Set					
Labs	MAE	MSE	R2	RMSE	MAPE
ALT	5.77	82.08	0.57	9.06	27.36
AST	4.58	53.73	0.49	7.33	20.26
Albumin	0.22	0.15	0.34	0.39	6.11

Test Set					
Labs	MAE	MSE	R2	RMSE	MAPE
ALT	5.92	86.81	0.59	9.32	28.04
AST	4.62	55.21	0.47	7.43	20.35
Albumin	0.22	0.15	0.32	0.39	6.21

Fig. 19: GB Regressor



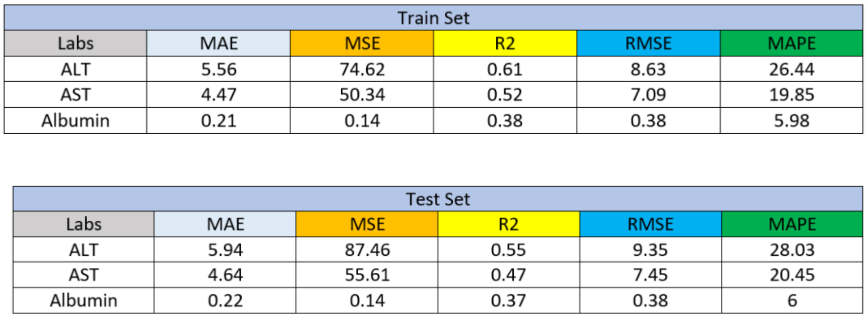
3 - RandomForest Regressor

Train Set					
Labs	MAE	MSE	R2	RMSE	MAPE
ALT	2.99	22.4	0.88	4.73	14.68
AST	2.12	11.66	0.89	3.42	9.55
Albumin	0.13	0.053	0.77	0.23	3.49

Train Set					
Labs	MAE	MSE	R2	RMSE	MAPE
ALT	6.5	100.87	0.48	10.04	31.13
AST	5.1	64	0.38	8	22.75
Albumin	0.25	0.18	0.18	0.42	6.75

Fig. 20: Random Forest Regressor

3 - Random Forest with Cross Validation and limited depth



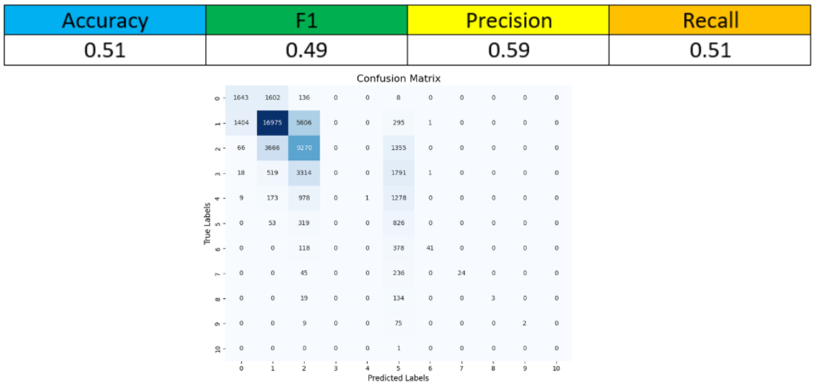
Train Set					
Labs	MAE	MSE	R2	RMSE	MAPE
ALT	5.56	74.62	0.61	8.63	26.44
AST	4.47	50.34	0.52	7.09	19.85
Albumin	0.21	0.14	0.38	0.38	5.98

Test Set					
Labs	MAE	MSE	R2	RMSE	MAPE
ALT	5.94	87.46	0.55	9.35	28.03
AST	4.64	55.61	0.47	7.45	20.45
Albumin	0.22	0.14	0.37	0.38	6

9

Fig. 21: Random Forest Regressor with limited depth and cross validation

5 - Decision Tree Classifier (10 classes)

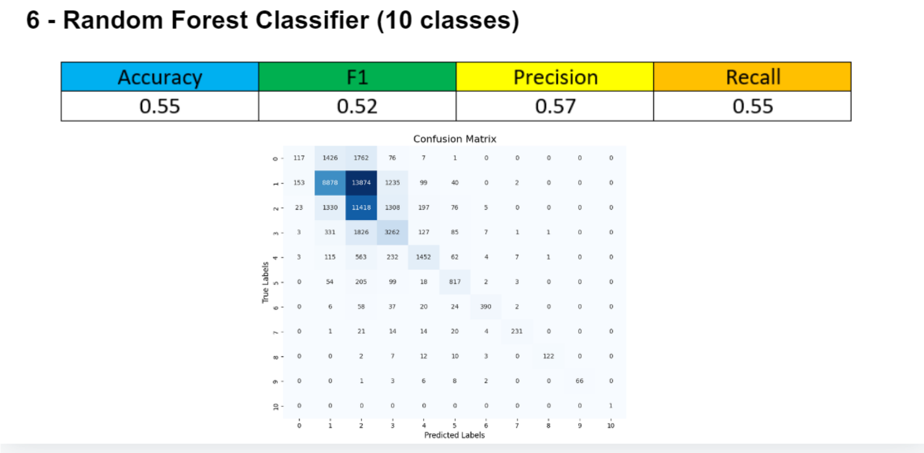


Accuracy	F1	Precision	Recall
0.51	0.49	0.59	0.51

Confusion Matrix

True Labels \ Predicted Labels	0	1	2	3	4	5	6	7	8	9	10
0	1643	1602	136	0	0	8	0	0	0	0	0
1	1404	16975	5606	0	0	295	1	0	0	0	0
2	66	3666	9270	0	0	1355	0	0	0	0	0
3	18	519	3314	0	0	1791	1	0	0	0	0
4	9	173	978	0	1	1278	0	0	0	0	0
5	0	53	319	0	0	826	0	0	0	0	0
6	0	0	118	0	0	378	41	0	0	0	0
7	0	0	45	0	0	236	0	24	0	0	0
8	0	0	19	0	0	134	0	0	3	0	0
9	0	0	9	0	0	75	0	0	0	2	0
10	0	0	0	0	0	1	0	0	0	0	0

Fig. 22: Decision Tree Classifier



6 - Random Forest Classifier (10 classes)

Accuracy	F1	Precision	Recall
0.55	0.52	0.57	0.55

Confusion Matrix

	0	1	2	3	4	5	6	7	8	9	10
0	117	1426	1762	76	7	1	0	0	0	0	0
1	153	8878	13874	1235	99	40	0	2	0	0	0
2	23	1330	11418	1308	197	76	5	0	0	0	0
3	3	331	1826	3262	127	85	7	1	1	0	0
4	3	115	563	232	1452	62	4	7	1	0	0
5	0	54	205	99	18	817	2	3	0	0	0
6	0	6	58	37	20	24	390	2	0	0	0
7	0	1	21	14	14	20	4	231	0	0	0
8	0	0	2	7	12	10	3	0	122	0	0
9	0	0	1	3	6	8	2	0	0	66	0
10	0	0	0	0	0	0	0	0	0	0	1

Fig. 23: Random Forest Classifier

8 Conclusion

This study highlights the effectiveness of machine learning models in predicting liver disease outcomes using a comprehensive dataset of over 500,000 patients. By employing various algorithms such as Random Forest, XGBoost, Support Vector Machine, Logistic Regression, and AdaBoost, we successfully identified key health indicators that significantly influence liver health, including ALT, AST, and albumin levels. Our findings demonstrate that the Random Forest model outperformed other algorithms, providing robust predictions while effectively handling complex and imbalanced health datasets.

The methodology encompassed meticulous data preprocessing, including filtering lab results, handling missing values, and conducting exploratory data analysis (EDA) to derive meaningful insights from patient demographics, vitals, and diagnoses. These steps were crucial in ensuring the quality and reliability of the data, thereby enhancing the performance of the predictive models. The integration of demographic factors further emphasized the need for personalized approaches to liver disease prediction.

The potential implications of this research extend beyond model accuracy. By facilitating earlier diagnoses and improving the personalization of treatment plans, machine learning could significantly contribute to better patient outcomes and reduced healthcare costs. Furthermore, the study lays the groundwork for future research aimed at incorporating additional variables, such as lifestyle factors and genetic predispositions, into the predictive models.

9 Future Work

Looking ahead, several avenues for future research can enhance the understanding and prediction of liver disease. First, expanding the dataset to include diverse populations and more granular demographic information will allow for the development of models that are more generalizable and applicable across different demographic groups. This inclusivity is critical in ensuring that predictions are accurate and relevant to all patient populations.

Second, integrating additional variables such as lifestyle choices (e.g., alcohol consumption, diet, exercise) and genetic markers could provide deeper insights into liver disease progression. These factors may significantly influence outcomes and can be instrumental in developing holistic predictive models.

Third, exploring advanced machine learning techniques, such as deep learning and ensemble methods that combine multiple models, may yield even higher predictive accuracy. These approaches could help in capturing intricate patterns and interactions within the data that traditional models might miss.

Finally, implementing these predictive models in clinical settings is essential. Collaborations with healthcare providers will facilitate the integration of machine learning tools into routine diagnostic processes, ultimately leading to improved early detection and management of liver diseases. Continuous evaluation and adaptation of the models based on real-world clinical feedback will also be crucial to ensure their relevance and effectiveness in patient care.