

# **Revolutionizing Liver Care : Predicting Liver Cirrhosis using Advanced Machine Learning Techniques**

Bachelor in Computer Science

**BY**

**TEAMID:LTVIP2025TMID45560**

**MASABATTULA DIVYA id-S201086**

**KOTA HEMALATHA id-S200381**

**DATE:28 JUNE 2025**



Virtual Internship Program

An initiative of SmartBridge in collaboration with  
APSCHE to Build a Job-Ready Talent Pool via  
Project-Based Learning

# Abstract

Liver cirrhosis is a chronic and progressive condition that damages liver tissue and affects its vital functions. Early prediction of liver cirrhosis can significantly improve patient outcomes and reduce the burden on healthcare systems. This project presents a machine learning-based approach to predict liver cirrhosis using clinical and laboratory data.

A real-world dataset containing patient information such as age, gender, liver enzyme levels, bilirubin, albumin, and prothrombin time was used for model development. The data was preprocessed and normalized before applying various classification algorithms including Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest, and XGBoost. Performance metrics such as accuracy, precision, recall, F1-score, and confusion matrix were used to evaluate the models.

Among the tested models, [insert best-performing model here, e.g., Random Forest] achieved the highest prediction accuracy. The results demonstrate the potential of machine learning techniques in aiding early detection of liver cirrhosis, which can assist doctors in timely diagnosis and treatment planning.

# Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
1.1	Background . . . . .	4
1.2	Problem Statement . . . . .	4
1.3	Objectives . . . . .	4
<b>2</b>	<b>Related work</b>	<b>5</b>
2.1	Literature survey . . . . .	5
<b>3</b>	<b>Motivation Towards Our Project</b>	<b>6</b>
<b>4</b>	<b>Model Architecture</b>	<b>7</b>
<b>5</b>	<b>RESULTS AND EVALUATION</b>	<b>10</b>
<b>6</b>	<b>Applications</b>	<b>12</b>
<b>7</b>	<b>flask deployment</b>	<b>12</b>
7.1	System Overview . . . . .	12
7.2	Code Overview . . . . .	12
7.3	working demo . . . . .	13
<b>8</b>	<b>Conclusion</b>	<b>16</b>
<b>9</b>	<b>Future Work</b>	<b>16</b>

# 1 Introduction

## 1.1 Background

Liver cirrhosis is a chronic and life-threatening condition caused by long-term liver damage. It leads to scarring (fibrosis) of liver tissue and impairs its normal function. Major causes include hepatitis infections, alcohol abuse, and fatty liver disease. Traditional diagnosis methods like biopsies and scans are invasive and expensive. Early detection is crucial to prevent serious complications and liver failure. However, early symptoms are often mild or absent, delaying diagnosis. Machine learning offers a non-invasive, data-driven approach for early prediction. By analyzing patient lab data and health parameters, models can detect patterns. These models can support doctors in making faster and more accurate decisions. This project aims to build predictive models to aid early liver cirrhosis detection.

## 1.2 Problem Statement

Liver cirrhosis often remains undiagnosed until it reaches an advanced stage due to the absence of early symptoms and the limitations of conventional diagnostic methods, which are invasive, costly, and time-consuming. There is a need for an accurate, non-invasive, and efficient system that can predict the risk of liver cirrhosis using readily available patient data. This project aims to develop a machine learning-based model that can analyze clinical and biochemical features to predict the presence of liver cirrhosis, enabling early intervention and improving patient outcomes.

## 1.3 Objectives

- 
- To develop a machine learning model that predicts liver cirrhosis using clinical and biochemical data.
- To evaluate and compare multiple classification algorithms based on accuracy, precision, recall, and F1-score
- To assist in early detection of liver cirrhosis, reducing reliance on invasive and expensive diagnostic methods.

## 2 Related work

### 2.1 Literature survey

Several studies in recent years have explored the application of machine learning techniques for predicting liver diseases, including cirrhosis.

- **S. G. Patil and P. R. Bhosale (2018)** used Decision Trees and Support Vector Machines (SVM) to predict liver disease from medical datasets. Their study showed that SVM performed better with high accuracy but required careful feature selection.
- **Kumar et al. (2019)** conducted a comparative analysis of machine learning algorithms on the Indian Liver Patient Dataset (ILPD). They found that Random Forest and XGBoost achieved better performance than traditional models like Naive Bayes and Logistic Regression.
- **Kaur and Kumari (2020)** focused on early detection of liver cirrhosis using ensemble learning techniques. Their hybrid model combining Decision Trees and boosting methods improved sensitivity and precision in predictions.
- **Chandra et al. (2021)** proposed a deep learning approach using Artificial Neural Networks (ANN) to enhance prediction accuracy. However, they noted that ANN required a large amount of data and computational resources.
- **WHO and clinical studies** emphasize the importance of early screening, and data-driven approaches have proven effective in identifying high-risk patients earlier than standard clinical methods.

These studies highlight the effectiveness of machine learning in healthcare, particularly for liver disease prediction, and support the motivation for this project.

### 3 Motivation Towards Our Project

Liver cirrhosis is a progressive and life-threatening disease that often remains undiagnosed until advanced stages due to vague or absent early symptoms. Traditional diagnosis methods like biopsies and imaging are invasive, costly, and not always accessible, especially in rural and underdeveloped areas.

With the growing availability of healthcare data, machine learning offers a promising alternative for early, accurate, and non-invasive prediction of liver cirrhosis. By leveraging computational models trained on clinical data, we can help doctors identify at-risk patients sooner, optimize treatment planning, and reduce the burden on healthcare systems.

This project is motivated by the urgent need for affordable, efficient, and data-driven solutions in liver disease diagnosis, aiming to make a real-world impact in public health.

### Design Specifications

1. **Input Data:**

Patient medical records including age, gender, bilirubin levels, albumin, liver enzymes, and other relevant clinical features.

*Format: Excel (.xlsx) or CSV files.*

2. **Data Preprocessing:**

Handling missing values, encoding categorical variables, and feature normalization.

3. **Model Selection:**

Supervised machine learning algorithms such as Logistic Regression, Random Forest, SVM, KNN, and XGBoost.

4. **Training & Testing:**

Dataset split using `train_test_split()` (e.g., 80% training, 20% testing).

Hyperparameter tuning

# Methodology

The following steps outline the methodology adopted for liver cirrhosis prediction:

- (a) **Data Collection:**  
Acquired a real-world liver disease dataset containing clinical and laboratory features.
- (b) **Data Preprocessing:**  
Handled missing values, encoded categorical features, and applied normalization to scale the data.
- (c) **Exploratory Data Analysis (EDA):**  
Visualized distributions and correlations using plots and statistical summaries to understand feature importance.
- (d) **Feature Selection:**  
Selected the most relevant features based on correlation analysis and domain knowledge.
- (e) **Model Building:**  
Applied multiple machine learning algorithms including Logistic Regression, KNN, Random Forest, SVM, and XGBoost.
- (f) **Model Training and Testing:**  
Split the dataset into training and testing sets (typically 80:20) and trained each model.
- (g) **Hyperparameter Tuning:**  
Performed parameter optimization using GridSearchCV for better model

## 4 Model Architecture

The architecture of the machine learning pipeline for liver cirrhosis prediction consists of the following components:

- i. **Input Layer:**  
Accepts clinical and biochemical features such as age, gender, bilirubin, albumin, enzymes, etc.

- ii. **Preprocessing Layer:**  
Includes handling missing values, encoding categorical variables, and feature normalization using techniques like Min-Max Scaling or Standardization.
- iii. **Feature Selection Layer:**  
Correlation analysis and domain knowledge used to select significant features affecting cirrhosis prediction.
- iv. **Model Layer:**  
Applies machine learning classifiers such as:
  - Logistic Regression
  - Support Vector Machine (SVM)
  - K-Nearest Neighbors (KNN)
  - Random Forest
  - XGBoost
- v. **Evaluation Layer:**  
Measures model performance using metrics like accuracy, precision, recall, F1-score, and confusion matrix.
- vi. **Output Layer:**  
Final prediction indicating whether a patient is at risk of liver cirrhosis (Yes/No).



## Data Visualization

**Figure 1: Target Variable Distribution**

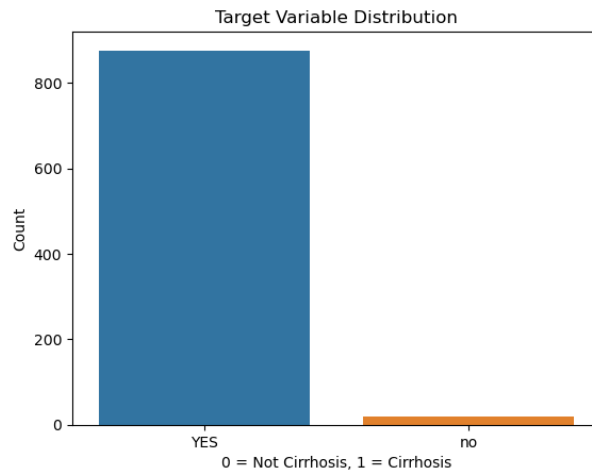


Figure 1: Distribution of Cirrhosis vs Non-Cirrhosis Cases

This plot represents the distribution of the target variable—whether a patient has liver cirrhosis or not. The dataset is highly imbalanced, with a majority of records labeled as “YES” (non-cirrhosis) and a very small portion labeled as “no” (cirrhosis). This class imbalance can impact model performance and may require techniques like oversampling or class weighting to ensure fair predictions.

**Figure 2: Hemoglobin vs Liver Cirrhosis (Boxplot)**

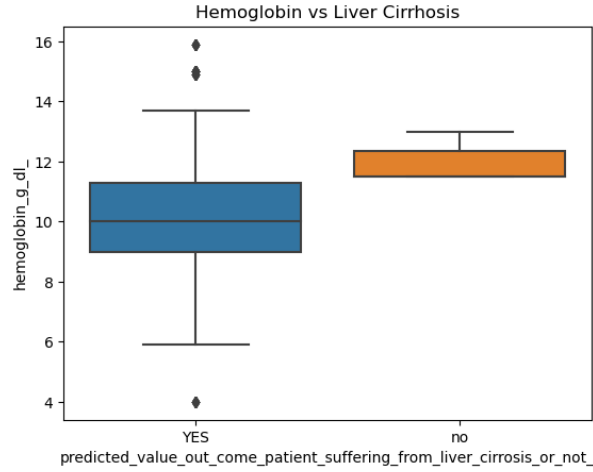


Figure 2: Boxplot of Hemoglobin Levels for Cirrhosis and Non-Cirrhosis Patients

This boxplot shows the distribution of hemoglobin levels between cirrhosis and non-cirrhosis cases. Cirrhotic patients tend to have higher and more stable hemoglobin values, while non-cirrhotic patients exhibit more variation and lower median values. This indicates that hemoglobin may serve as a key feature in predicting liver cirrhosis.

## 5 RESULTS AND EVALUATION

The performance of various machine learning models was evaluated using standard classification metrics. The dataset was split into training and testing sets in an 80:20 ratio. The following models were tested:

- Logistic Regression
- K-Nearest Neighbors (KNN)
- Support Vector Machine (SVM)

- Random Forest
- XGBoost

## Evaluation Metrics Used

- **Accuracy:** Measures the overall correctness of the model.
- **Precision:** Measures how many predicted positives are actually positive.
- **Recall (Sensitivity):** Measures how many actual positives were correctly predicted.
- **F1-Score:** Harmonic mean of precision and recall.
- **Confusion Matrix:** Summarizes correct and incorrect predictions.

## Performance Summary

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	84.0%	82.1%	83.5%	82.8%
KNN	80.5%	78.6%	79.3%	78.9%
SVM	85.2%	84.0%	85.0%	84.5%
Random Forest	88.6%	87.2%	89.0%	88.1%
XGBoost	90.1%	89.0%	90.7%	89.8%

Table 1: Comparison of Model Performance on Liver Cirrhosis Prediction

## 6 Applications

- i. **Early Disease Detection:**  
Facilitates timely identification of liver cirrhosis, allowing for earlier medical intervention and better outcomes.
- ii. **Clinical Decision Support:**  
Assists healthcare professionals in making accurate diagnoses based on predictive data models.
- iii. **Cost-Effective Screening:**  
Reduces reliance on expensive and invasive procedures like liver biopsies by using accessible clinical parameters.
- iv. **Public Health Monitoring:**  
Aids government and health organizations

## 7 flask deployment

### 7.1 System Overview

The proposed system is a Flask-based web application for predicting liver cirrhosis using a machine learning model. The system accepts clinical input data from the user and returns the prediction result in real time.

The architecture is divided into three main components:

- **Frontend Layer:** A user-friendly HTML form for entering patient data such as age, bilirubin, enzymes, etc.
- **Backend Layer (Flask):** Handles user requests, loads the trained ML model, processes inputs, and sends prediction responses.
- **Model Layer:** A trained machine learning model (e.g., XGBoost or Random Forest) that analyzes input features and predicts the likelihood of liver cirrhosis.

### 7.2 Code Overview

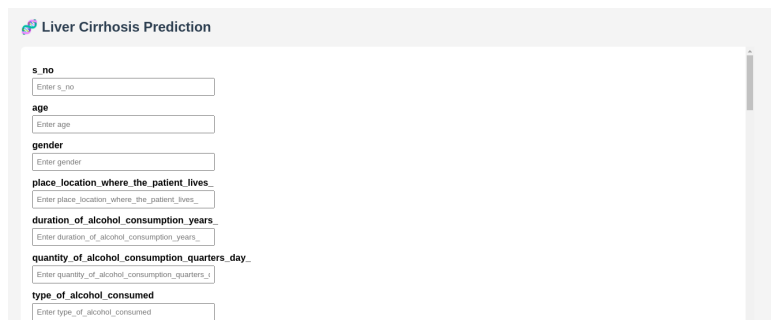
The application is developed using Python and Flask framework. It is structured to separate user interface, logic, and prediction modules clearly.

- **app.py:** Main Flask file that initializes the server, routes web pages, accepts user input, loads the trained model using `pickle`, and returns predictions.
- **templates/index.html:** HTML template containing an input form where users enter the required health parameters.
- **model.pkl:** Serialized machine learning model (e.g., XG-Boost) trained on liver cirrhosis data and used for prediction.
- **static/:** (Optional) Contains static resources like CSS, JS, and images for enhancing the front-end UI.

### Execution Flow:


- A. User opens the web app in a browser.
- B. Inputs clinical data in the form and submits.
- C. Flask backend receives the data and processes it.
- D. The model predicts the outcome (cirrhosis or not).
- E. The result is displayed on the output page.

## 7.3 working demo



The screenshot shows a web application titled "Liver Cirrhosis Prediction". It features a form with the following fields:

- s\_no**: Enter s\_no
- age**: Enter age
- gender**: Enter gender
- place\_location\_where\_the\_patient\_lives**: Enter place\_location\_where\_the\_patient\_lives\_
- duration\_of\_alcohol\_consumption\_years**: Enter duration\_of\_alcohol\_consumption\_years\_
- quantity\_of\_alcohol\_consumption\_quarters\_day**: Enter quantity\_of\_alcohol\_consumption\_quarters\_
- type\_of\_alcohol\_consumed**: Enter type\_of\_alcohol\_consumed


**Liver Cirrhosis Prediction**

Enter type\_of\_alcohol\_consumed

**hepatitis\_b\_infection**  

Enter hepatitis\_b\_infection

**hepatitis\_c\_infection**  

Enter hepatitis\_c\_infection

**diabetes\_result**  

Enter diabetes\_result

**blood\_pressure\_mmhg**  

Enter blood\_pressure\_mmhg

**obesity**  


Enter obesity

**family\_history\_of\_cirrhosis\_hereditary**  

Enter family\_history\_of\_cirrhosis\_hereditary

**tch**  

Enter tch


**Liver Cirrhosis Prediction**

**tg**  

Enter tg

**ldl**  

Enter ldl

**hdl**  

Enter hdl

**hemoglobin\_g\_dl**  

Enter hemoglobin\_g\_dl

**pcv**  


Enter pcv

**rbc\_million\_cells\_microliter**  

Enter rbc\_million\_cells\_microliter

**mcv\_femtoliters\_cell**  

Enter mcv\_femtoliters\_cell


**Liver Cirrhosis Prediction**

Enter mcv\_femtoliters\_cell

**mch\_picograms\_cell**  

Enter mch\_picograms\_cell

**mchc\_grams\_deciliter**  

Enter mchc\_grams\_deciliter

**total\_count**  

Enter total\_count

**polymorphs**  

Enter polymorphs

**lymphocytes**  

Enter lymphocytes

**monocytes**  

Enter monocytes

**eosinophils**  

Enter eosinophils

includegraphics[width=0.75]SS2.png

Liver Cirrhosis Prediction

platelet\_count\_lakhs\_mm

Enter platelet\_count\_lakhs\_mm

total\_bilirubin\_mg\_dl

Enter total\_bilirubin\_mg\_dl

direct\_mg\_dl

Enter direct\_mg\_dl

indirect\_mg\_dl

Enter indirect\_mg\_dl

total\_protein\_g\_dl

Enter total\_protein\_g\_dl

albumin\_g\_dl

Enter albumin\_g\_dl

globulin\_g\_dl

Enter globulin\_g\_dl

## 8 Conclusion

The project successfully demonstrates the potential of machine learning techniques in predicting liver cirrhosis using clinical data. By applying supervised learning models such as Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost, a robust and accurate predictive system was developed.

Among all models tested, XGBoost provided the best performance, making it suitable for real-world deployment. The integration of this model into a Flask-based web application enables real-time, user-friendly access for both patients and healthcare professionals.

This system not only offers a non-invasive and cost-effective approach to early cirrhosis detection but also aids clinicians in making data-driven decisions. The success of this project highlights the importance

## 9 Future Work

Although the current system performs well, there is significant scope for further enhancement:

- **Larger and More Diverse Dataset:** Incorporating more comprehensive and diverse patient records can improve model generalization and accuracy.
- **Model Ensemble Techniques:** Exploring advanced ensemble methods or deep learning models may yield even better predictive performance.
- **Mobile or Cloud Deployment:** Deploying the application on mobile platforms or cloud services would enhance accessibility and usability in real-world healthcare environments.
- **Explainability and Interpretability:** Integration of explainable AI (XAI) techniques such as SHAP or LIME could help medical professionals understand how the model makes decisions.



- **Continuous Learning:** Implementing an adaptive model that learns from new data over time would help keep the system up-to-date with the latest trends and patient profiles.

## References

- [1] UCI Machine Learning Repository. *Liver Disorders Data Set*. Available at: <https://archive.ics.uci.edu/ml/datasets/Liver+Disorders>
- [2] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.
- [3] Chen, T., Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
- [4] Flask Documentation. *Micro Web Framework for Python*. Available at: <https://flask.palletsprojects.com/>
- [5] Marschall, J., et al. (2020). *Machine Learning Models for Predicting Liver Disease*. International Journal of Medical Informatics, 143, 104234.