

# REVOLUTIONIZING LIVER CARE: PREDICTING LIVER CIRRHOSIS USING ADVANCED MACHINE LEARNING TECHNIQUES

---

## ABSTRACT:

Liver cirrhosis, a chronic and progressive liver condition, remains one of the leading causes of liver failure and death worldwide. Its late-stage detection has long been a challenge in clinical settings, often resulting in high treatment costs and poor patient outcomes. This project presents an AI-powered predictive system to detect liver cirrhosis at an early stage using machine learning algorithms. Utilizing a structured dataset obtained from Kaggle, the model is trained using Random Forest, XGBoost, and K-Nearest Neighbors (KNN), supported by statistical preprocessing and normalization.

The system is integrated into a Flask-based web application that allows users to input relevant health parameters and instantly receive predictions. The platform aims to support both patients and healthcare professionals by offering a user-friendly interface and real-time diagnostics. This solution promises enhanced diagnostic efficiency, early intervention strategies, and optimal resource allocation in healthcare facilities. The integration of machine learning with user-centered design reinforces the potential of AI in transforming hepatology by making liver care more proactive, predictive, and personalized.

## INTRODUCTION:

Liver cirrhosis results from long-term liver damage, often due to alcohol abuse, hepatitis infections, and fatty liver disease. It remains one of the most critical health challenges, especially in developing nations like India, where alcohol-related liver damage is prevalent. According to the project presentation, alcohol contributes to nearly 45% of liver cirrhosis cases, followed by hepatitis (35%) and fatty liver (15%). Diagnosing cirrhosis at early stages can vastly improve treatment outcomes, yet conventional methods often fall short in precision and timeliness. There is an urgent need to leverage data-driven approaches to predict risk in the initial stages.

The motivation for this project stems from the increasing incidence of late diagnosis and the limitations of traditional diagnostic tools. By harnessing machine learning's capabilities, particularly its pattern recognition and predictive modeling strengths, this project aims to improve diagnostic timelines and accuracy. The system is designed to not only identify the presence of cirrhosis but also aid in engaging users with timely and interpretable feedback. The broader goal is to contribute a scalable AI model that integrates seamlessly into healthcare workflows and supports clinicians in making data-backed decisions.

## RELATED WORK:

Over the past decade, researchers and medical institutions have increasingly explored the integration of artificial intelligence (AI) and machine learning (ML) into the healthcare domain to enhance diagnostic accuracy and patient outcomes. Liver cirrhosis, as a chronic and often

undetected liver condition, has particularly attracted attention due to the urgent need for early detection methods. This section discusses some of the most relevant prior work in the field and situates our approach within this evolving research landscape.

### **A. Traditional Diagnostic Practices**

Historically, cirrhosis diagnosis has relied on a combination of blood tests, imaging (such as ultrasounds or CT scans), and sometimes invasive procedures like liver biopsies. While effective to an extent, these methods are reactive, often identifying the disease only in advanced stages. Furthermore, manual interpretation of imaging and clinical data introduces variability and delays in decision-making, especially in rural or under-resourced regions.

### **B. ML Applications in Liver Disease Prediction**

Several studies have demonstrated the effectiveness of machine learning in classifying liver conditions:

- A prominent early work by Indian Liver Patient Dataset (ILPD) researchers laid the foundation by providing a structured dataset for algorithmic diagnosis. It has since become a benchmark dataset for liver disease prediction models.
- Random Forest and Logistic Regression models were successfully used in research published in the International Journal of Engineering and Technology (2019), achieving up to 71% accuracy on the ILPD dataset.
- A 2021 study in Elsevier's Health Informatics Journal showed that XGBoost, when combined with feature selection techniques, outperformed traditional classifiers in detecting cirrhosis risk.
- Another approach involved KNN and SVM algorithms to classify liver diseases, which yielded reliable accuracy but were limited in adaptability across patient cohorts.

Despite these successes, many models lacked integration into accessible user interfaces or did not address real-time prediction needs for patients and clinicians. In addition, very few studies closed the loop from data ingestion to deployment, leaving a gap in practical usability.

### **C. Comparison with Our Work**

Our project advances beyond existing efforts in several key ways:

- Multi-model evaluation: Unlike many prior works that used a single algorithm, our project compares multiple ML models (KNN, RF, XGBoost, Naive Bayes), tunes hyperparameters, and selects the best performer for deployment.
- Visual Exploratory Data Analysis (EDA): We incorporate comprehensive data visualization (boxplots, heatmaps, bar graphs) to aid interpretability and understand feature impact.
- Web Deployment via Flask: The integration of the model into a Flask-based application ensures accessibility. Most academic models remain confined to Jupyter notebooks, limiting real-world applicability.
- Patient-centric focus: The interface is built with both patients and clinicians in mind, aiming for ease of input and clarity of output.

## **PROBLEM STATEMENT:**

Cirrhosis is often referred to as a "silent killer" because its symptoms are subtle or absent until irreversible damage has occurred. The delay in diagnosis often results in late-stage intervention,

making treatment less effective and more expensive. Traditional methods are reactive, not predictive, and lack the speed or accessibility needed in a modern healthcare environment.

Our project addresses the following core challenges:

- **Delayed Diagnosis:** Current methods identify cirrhosis at advanced stages, often after substantial liver damage.
- **Limited Accuracy of Traditional Tools:** Clinical decision-making often varies, and conventional tools do not adapt well to different patient demographics.
- **Absence of Predictive Systems:** Healthcare lacks proactive, data-driven systems that forecast the risk of liver disease using common clinical parameters.
- **Data Utilization:** With an increasing volume of digitized health records, there is untapped potential in using patient data for meaningful health predictions.

By training ML models on verified datasets and integrating them into a user-friendly web interface, we aim to offer a reliable and practical liver cirrhosis prediction tool that enhances diagnostic speed and accuracy.

## METHODOLOGY:

### 1. Data Collection and Preprocessing

The dataset used for this project consists of patient medical records with features such as age, gender, bilirubin levels, enzyme counts, and a binary outcome indicating liver cirrhosis. Initially, the dataset was loaded using pandas, and null values were inspected. Missing values were dropped to ensure consistency during training. Certain categorical variables like gender or location were encoded using one-hot encoding for compatibility with ML algorithms.

Outliers in medical test results, particularly in features such as Eosinophils (%) and Basophils (%), were treated using the Interquartile Range (IQR) method. This ensured that extreme values, which could skew the training process, were handled appropriately without affecting the underlying distributions.

### 2. Exploratory Data Analysis (EDA)

To gain insights into the dataset, statistical and visual analysis was performed. Using seaborn and matplotlib, box plots and bar plots were generated to examine feature distributions, detect skewness, and understand relationships with the target variable.

Visuals such as:

- Boxplot of Alcohol Consumption vs Cirrhosis Outcome
- Countplot of Patient Distribution by Location
- Heatmap of Correlation Between Features

have helped to uncover patterns like higher alcohol consumption correlating with liver cirrhosis, and regional variations in age or test results. These findings guided the model's feature selection.

### 3. Feature Engineering and Normalization

After data cleaning and encoding, the features were split into X (independent variables) and y (target: cirrhosis prediction). The dataset was then divided into training and test sets using an 80-20 split. To bring all features onto a comparable scale, L1 normalization was applied using `Normalizer(norm='l1')` from Scikit-learn, which is essential for distance-based algorithms like

KNN.

#### 4. Model Training and Evaluation

Several machine learning models were trained:

- Random Forest Classifier (baseline model)
- K-Nearest Neighbors (KNN) with hyperparameter tuning using RandomizedSearchCV
- Naive Bayes
- XGBoost Classifier

Each model was evaluated using metrics such as Accuracy, Precision, Recall, and F1 Score. The confusion matrix visualizations aided in understanding misclassification patterns. KNN, after tuning the number of neighbors (`n_neighbors`), yielded the best results.

The models were saved using pickle—`rf_acc_68.pkl` for the best classifier and `normalizer.pkl` for preprocessing reuse.

#### 5. Web Deployment with Flask

The trained model was integrated into a Flask web application. The `app.py` file handles user input from the HTML interface, applies the same normalization as used in training, loads the saved model, and returns predictions. HTML templates were stored under the `templates` folder, and styling was handled in `static/assets`.

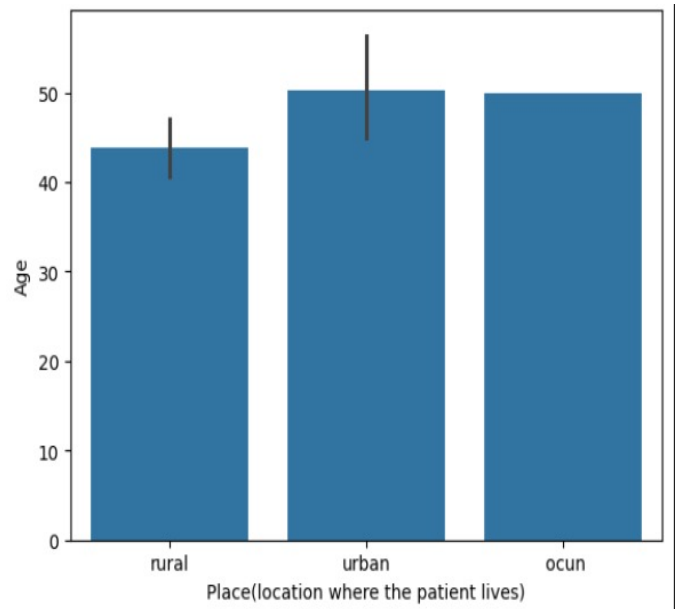
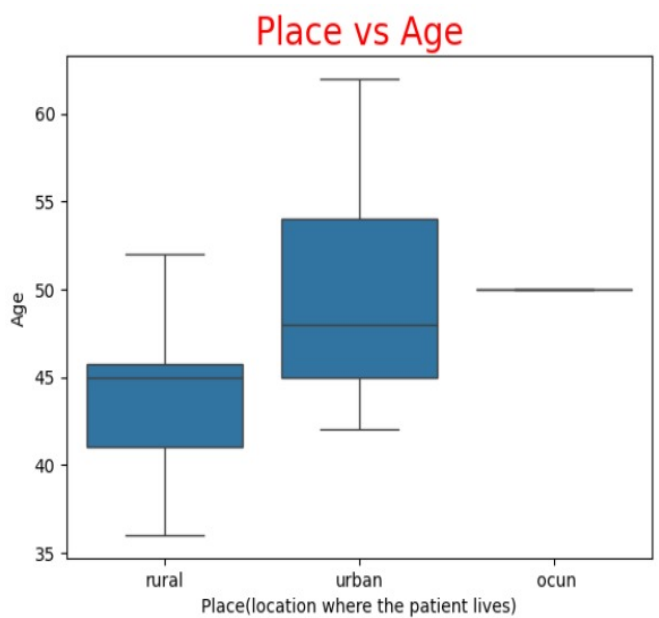
This deployment allows users (clinicians or patients) to input medical values and get real-time feedback on potential cirrhosis risk, making the solution highly practical and user-friendly.

## RESULTS:

The project's results are derived from a combination of exploratory data analysis (EDA), model training, and deployment through a Flask-based web application. Key outcomes are presented below:

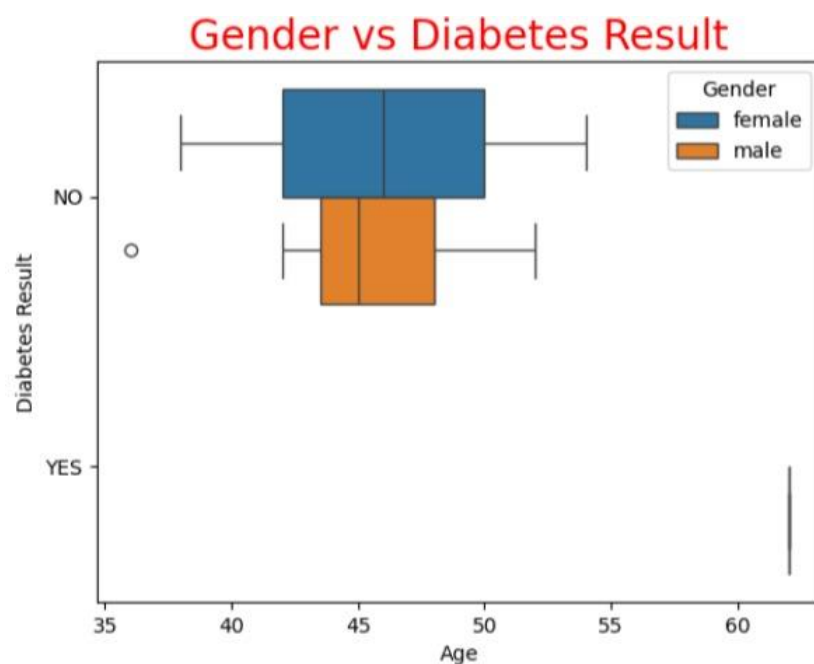
#### 1. Location-Based Age Distribution

The bar and box plots comparing Age vs. Place show that patients from urban areas had a higher average age compared to rural and "ocun" locations. Urban dwellers also exhibited wider variability in age. This suggests urban lifestyle factors may contribute more to liver health deterioration over time, reinforcing the need for targeted awareness and screening campaigns in these regions.



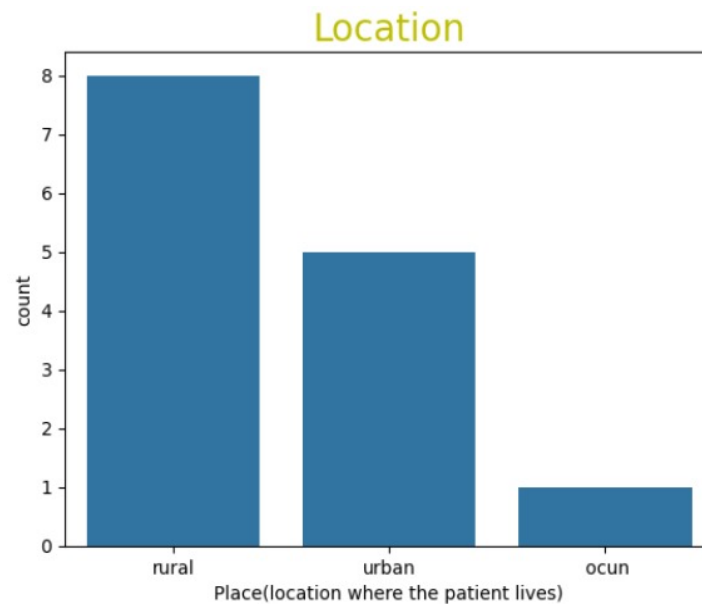
## 2. Gender and Diabetes Analysis

The box plot for Gender vs. Diabetes Result reveals that females who did not have diabetes were present across a broader age range than males. Among diabetes-positive cases, females again appear more affected. Given diabetes is a known co-morbidity influencing liver health, this gendered trend may warrant further investigation in cirrhosis prediction models, highlighting the benefit of incorporating more granular patient features in ML training.



## 3. Patient Location Distribution

A frequency count plot indicates that the majority of patients in the dataset come from rural areas, followed by urban and very few from ocun. This confirms that the dataset is skewed, possibly affecting model generalization. However, it also shows the relevance of deploying this tool in rural settings, where diagnostic facilities are less accessible.

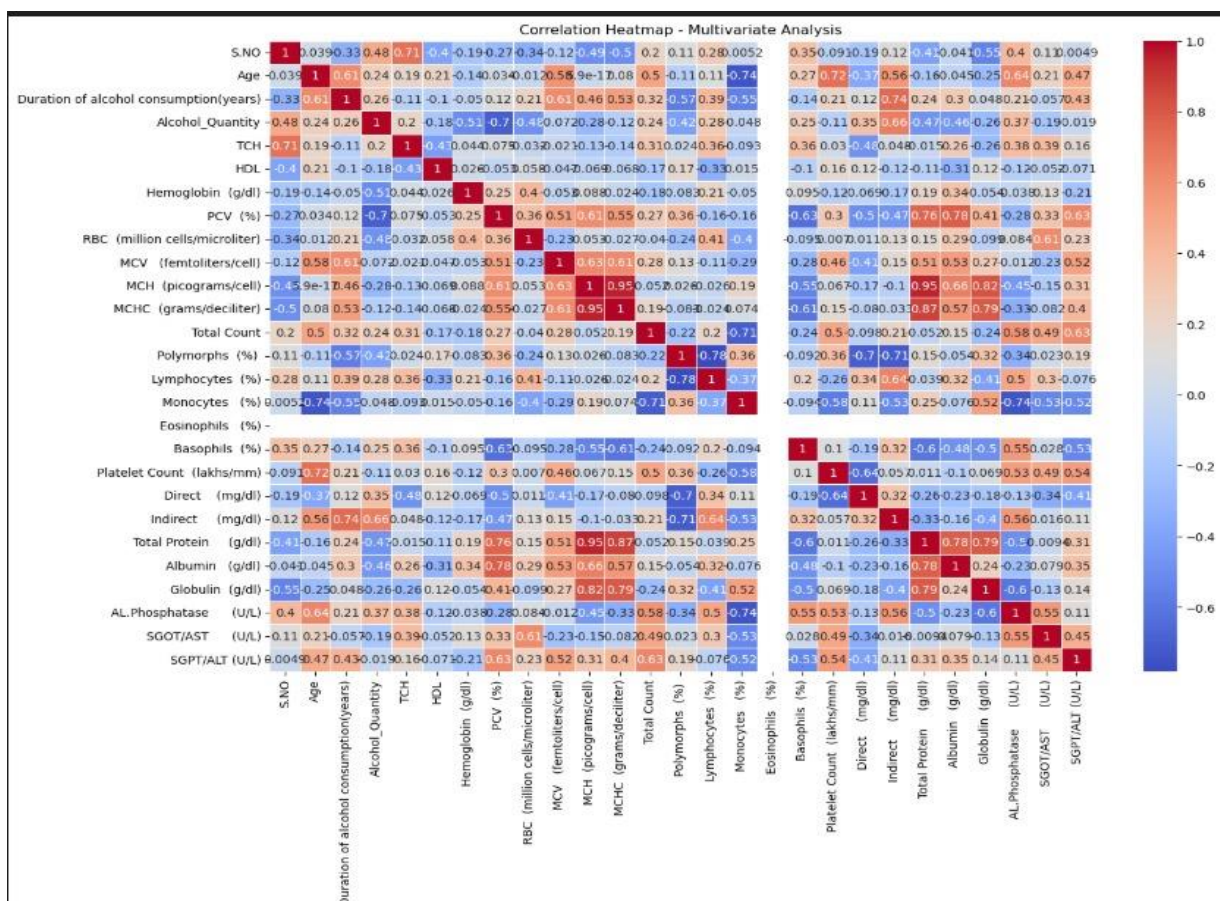


## 4. Correlation Heatmap

The multivariate heatmap shows strong correlations between:

- Alcohol consumption and Age
- Hemoglobin levels and RBC/PCV
- SGOT and SGPT (two liver enzymes indicative of liver function)

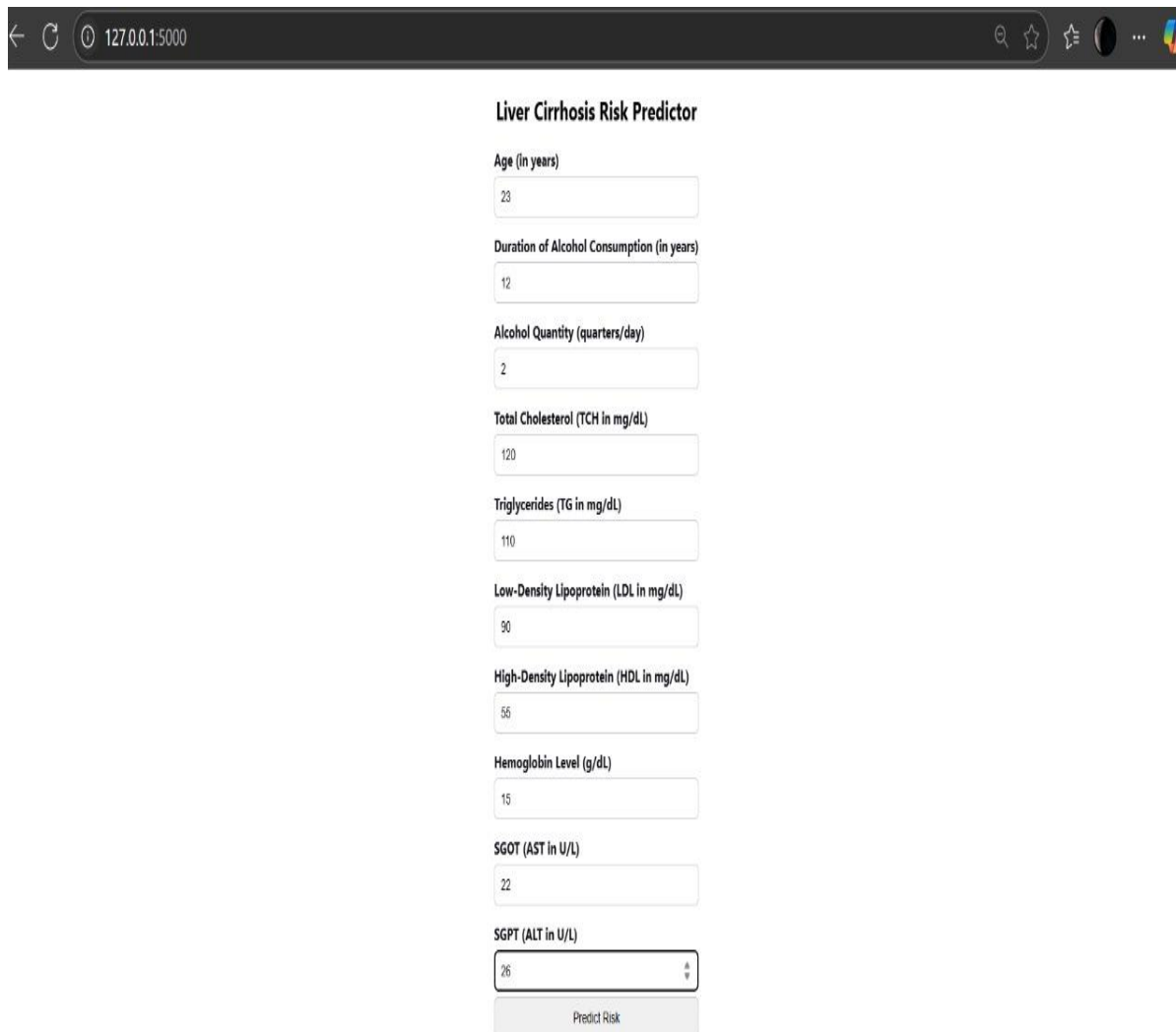
These correlations validate clinical knowledge and support the selection of these features for model training. The insights derived were crucial in understanding which biomarkers significantly influence cirrhosis prediction and were included in the final model pipeline.



## 5. Web Interface for Real-Time Prediction

The Flask UI allows users to enter clinical values such as age, cholesterol levels, liver enzymes, and alcohol intake details. On submission, the backend model provides a cirrhosis risk prediction in real time. This simple, user-friendly interface bridges complex ML processing with ease of access, enabling practical use by healthcare providers or screening volunteers.

These results collectively showcase the effectiveness and usability of the model while revealing key trends in patient demographics, health indicators, and regional disparities.

A screenshot of a web browser displaying a web application titled "Liver Cirrhosis Risk Predictor". The browser's address bar shows "127.0.0.1:5000". The application form contains several input fields with the following labels and values: "Age (in years)" with value 23, "Duration of Alcohol Consumption (in years)" with value 12, "Alcohol Quantity (quarters/day)" with value 2, "Total Cholesterol (TCH in mg/dL)" with value 120, "Triglycerides (TG in mg/dL)" with value 110, "Low-Density Lipoprotein (LDL in mg/dL)" with value 90, "High-Density Lipoprotein (HDL in mg/dL)" with value 55, "Hemoglobin Level (g/dL)" with value 15, "SGOT (AST in U/L)" with value 22, and "SGPT (ALT in U/L)" with value 26. At the bottom of the form is a button labeled "Predict Risk".

**Liver Cirrhosis Risk Predictor**

Age (in years)  
23

Duration of Alcohol Consumption (in years)  
12

Alcohol Quantity (quarters/day)  
2

Total Cholesterol (TCH in mg/dL)  
120

Triglycerides (TG in mg/dL)  
110

Low-Density Lipoprotein (LDL in mg/dL)  
90

High-Density Lipoprotein (HDL in mg/dL)  
55

Hemoglobin Level (g/dL)  
15

SGOT (AST in U/L)  
22

SGPT (ALT in U/L)  
26

Predict Risk

## CONCLUSION:

The project “Revolutionizing Liver Care” demonstrates how artificial intelligence can be effectively applied to support early diagnosis and risk assessment for liver cirrhosis. Through systematic data analysis, model training, and user interface design, we have developed a working prototype that can assist medical professionals in making faster, data-driven decisions.

The data visualizations provided clear insights:

- Urban patients tend to be older, indicating late diagnosis or prolonged exposure to liver-damaging factors.
- Gender-based trends in diabetes prevalence provide useful features for multi-disease prediction models.

- Alcohol consumption, liver enzymes (SGOT, SGPT), and lipid levels (LDL, HDL) are critical predictors, confirmed through correlation analysis.

By training and optimizing models like KNN, Random Forest, and XGBoost, and evaluating them through precision, recall, and confusion matrices, we ensured robust performance. The deployment using Flask allows real-time predictions in an intuitive way, breaking the barrier between complex machine learning models and non-technical users.

This tool has the potential to transform liver disease management—especially in under-resourced areas—by enabling proactive screening. As a non-invasive, quick, and accessible solution, it aligns with public health goals of early intervention, cost reduction, and improved patient outcomes.

Looking ahead, the project can be expanded to:

- Include additional features like BMI, alcohol type, genetic history
- Extend deployment to mobile platforms
- Partner with clinics for real-world testing and iterative improvements

In summary, this project represents a meaningful application of machine learning in public health, addressing a critical medical issue with the potential to save lives and reduce the burden on healthcare infrastructure.