PROJECT REPORT ON

"Revolutionizing Liver Care-Predicting Liver Cirrhosis Using Advanced Machine Learning Techniques"

BY: Swaraj Patil

Prisha Kumar

Prajakta Padwalkar

Atharva Chikane

INDEX

INTRODUCTION

Project Overview

Liver cirrhosis is a progressive and life-threatening condition caused by prolonged liver damage, leading to scarring (fibrosis) and impaired liver function. It can result from chronic hepatitis, excessive alcohol consumption, fatty liver disease, autoimmune disorders, and other factors. If left undiagnosed or untreated, cirrhosis can lead to severe complications such as liver failure, portal hypertension, and liver cancer. Early detection is critical for effective intervention, yet traditional diagnostic methods, such as liver biopsies and imaging scans, can be invasive, costly, and resource-intensive.

This project utilizes advanced machine learning techniques to develop a predictive model that analyses various patient data, including medical history, lab results, imaging findings, and lifestyle factors. By identifying complex patterns and correlations within these datasets, the model aims to provide an accurate assessment of a patient's likelihood of developing or progressing into cirrhosis.

The predictive system will be trained and validated using real-world clinical data to ensure accuracy and reliability. It will then be integrated into a user-friendly web application, allowing healthcare professionals to input patient data and receive real-time risk assessments. This tool will support early diagnosis, enabling timely medical interventions and personalized treatment strategies.

By leveraging artificial intelligence for predictive healthcare, this project aims to revolutionize liver disease management, offering a non-invasive, efficient, and scalable approach to detecting cirrhosis and improving patient outcomes.

Purpose

The purpose of this project is to develop an AI-driven predictive model that assists in the early detection and assessment of liver cirrhosis, enabling timely medical intervention and improved patient care. Liver cirrhosis is a progressive condition that can lead to severe complications such as liver failure and cancer if not diagnosed and managed early. Traditional diagnostic methods, including biopsies and imaging, can be invasive, expensive, and require specialized expertise, making early detection challenging.

By leveraging machine learning techniques, this project aims to analyse diverse patient data—such as medical history, laboratory test results, imaging scans, and lifestyle factors—to identify key patterns associated with cirrhosis. The predictive model will provide healthcare professionals with a non-invasive, data-driven tool for assessing cirrhosis risk, aiding in early diagnosis, personalized treatment planning, and better disease management.

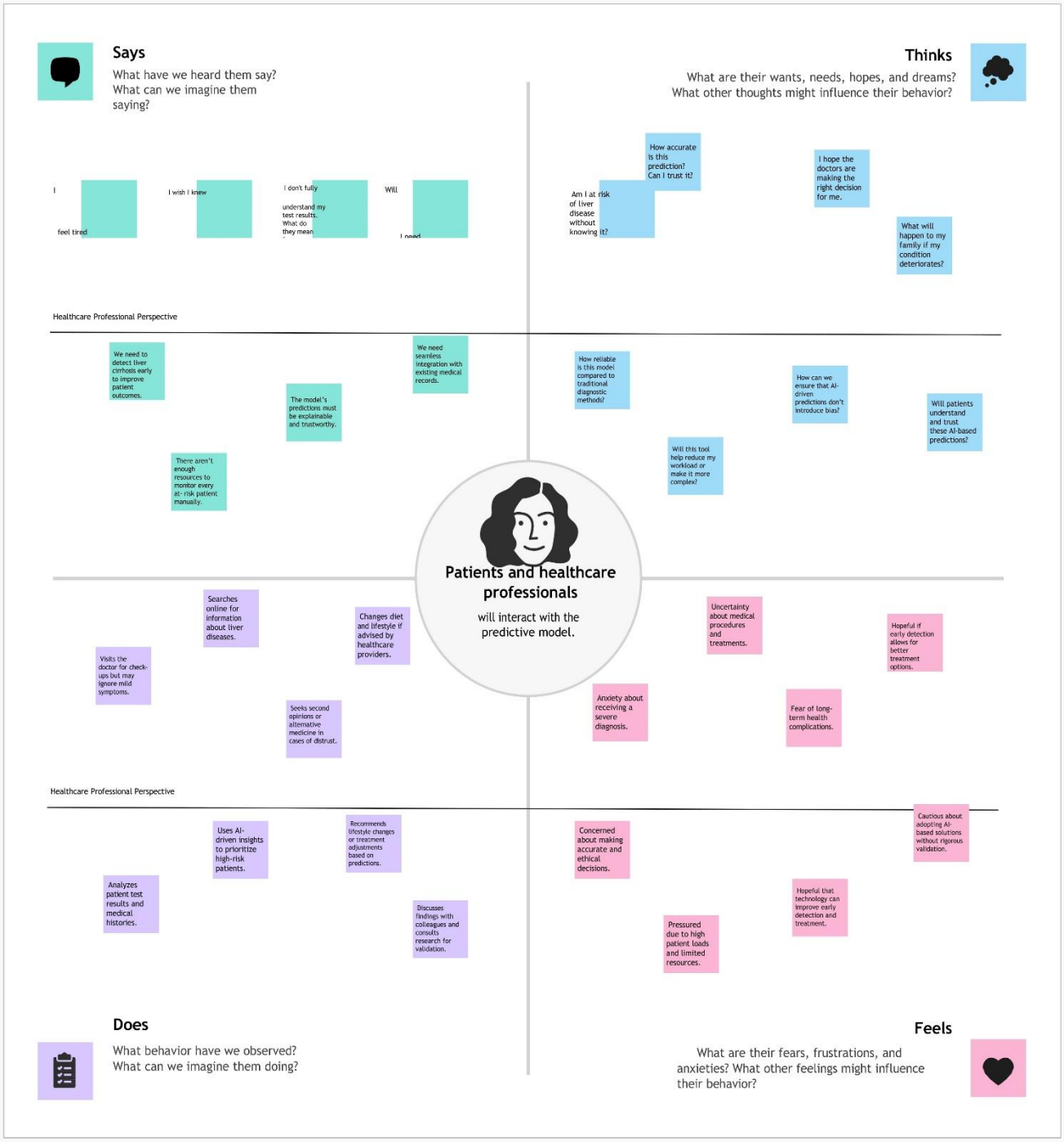
Furthermore, this project seeks to bridge the gap between AI and healthcare by integrating the model into a user-friendly web application. This will allow clinicians to input patient data and receive real-time risk assessments, enhancing decision-making and improving patient outcomes. Ultimately, the goal is to make liver disease prediction more accessible, efficient, and reliable, reducing the burden on healthcare systems and enhancing preventive care strategies.

IDEATHON PHASE

Problem Statement

Liver cirrhosis is a severe and progressive condition caused by long-term liver damage, leading to scarring and impaired function. It is often the result of chronic liver diseases such as hepatitis, fatty liver disease, and prolonged alcohol consumption. If left undiagnosed or untreated, cirrhosis can lead to life-threatening complications, including liver failure and an increased risk of liver cancer. Early detection is crucial to prevent irreversible damage and improve patient outcomes. This project aims to develop a predictive model using advanced machine learning techniques to assess the likelihood of liver cirrhosis in patients. By analysing a range of patient data, including medical history, laboratory test results, imaging scans, and lifestyle factors, the model will provide valuable insights for healthcare professionals. The integration of this predictive system into healthcare frameworks will assist in early diagnosis, proactive treatment planning, and efficient resource allocation, ultimately improving disease management and patient care. The primary purpose of this project is to enhance liver disease diagnosis and management through predictive analytics. By leveraging machine learning, the model will help identify high-risk patients, support personalized treatment planning, and optimize healthcare resources. Early prediction of cirrhosis progression will allow medical professionals to implement timely interventions, adjust treatment strategies, and provide targeted lifestyle recommendations to slow disease progression. Furthermore, integrating this predictive model into healthcare systems can improve clinical decision-making and streamline patient management, reducing the burden on medical facilities.

Empathy Map Canvas



Brainstorming

1.Problem Identification

- Liver cirrhosis is often diagnosed too late, leading to severe health consequences.
- High-risk individuals may go undetected due to inefficient screening processes.
- Manual diagnosis is time-consuming and prone to human error.
- Healthcare systems lack predictive tools for early intervention.

2. Key Stakeholders

- Doctors & Healthcare Providers – Need accurate, fast, and explainable predictions.
- Patients – Need early diagnosis and clear treatment guidance.
- Hospitals & Clinics – Require efficient and scalable AI-driven healthcare tools.
- Medical Researchers – Seek AI advancements in disease prediction.

3. Machine Learning-Based Prediction Model

- The system will analyse patient medical records, including lab test results, medical history, and lifestyle factors, to predict the likelihood of liver cirrhosis.
- A classification model (e.g., Random Forest, XGBoost, or Deep Learning) will categorize patients into low-risk, moderate-risk, or high-risk groups.
- Model explainability tools (e.g., SHAP, LIME) will be used to provide transparency on feature importance.

- The AI model will continuously learn from new patient data through automated retraining pipelines.

4. Potential Challenges & Risks

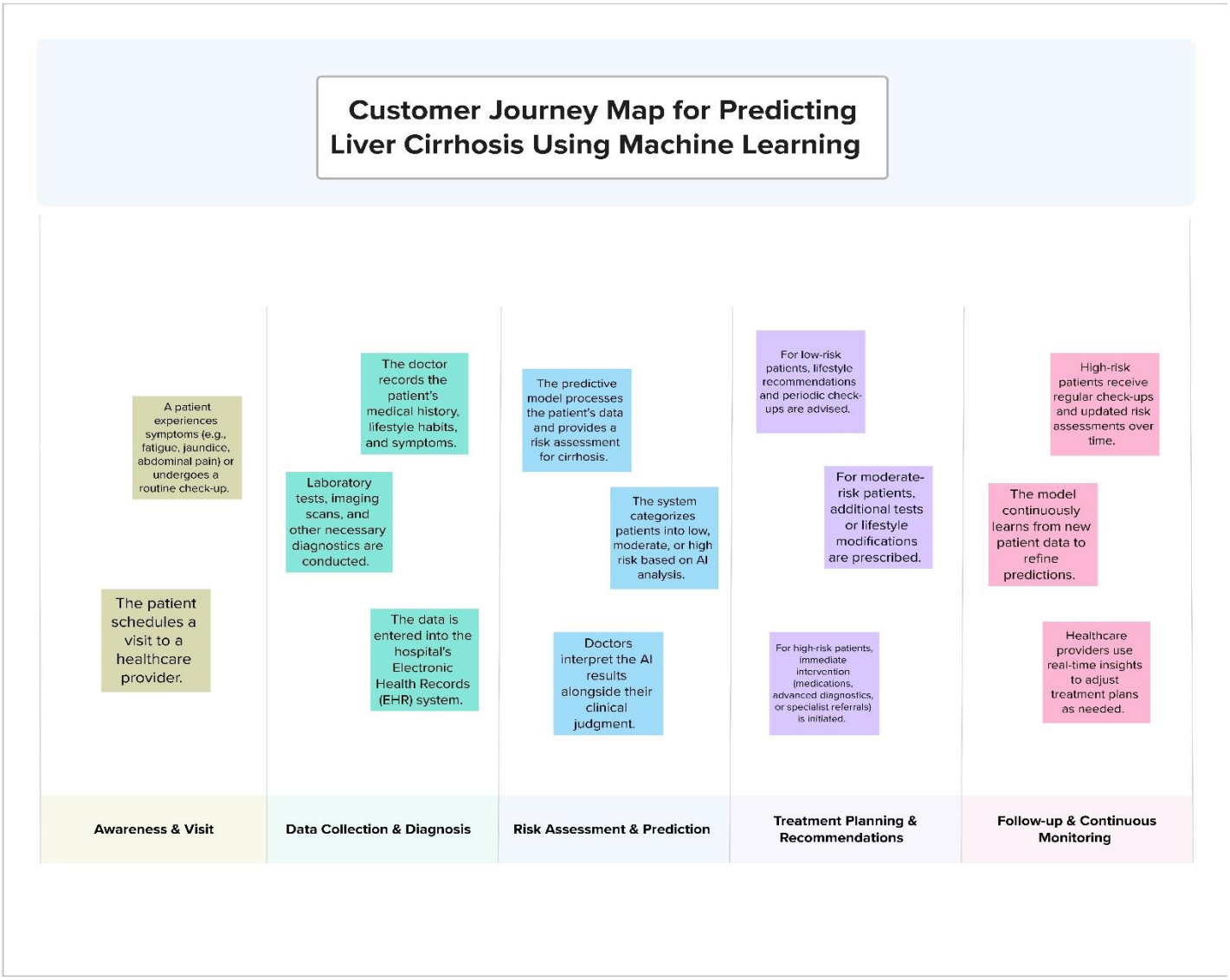
- Data Availability & Privacy – Ensuring access to sufficient patient records while maintaining confidentiality.
- Bias in AI Models – Preventing discrimination against specific demographics.
- Integration with Healthcare Systems – Making the AI tool compatible with existing hospital infrastructure.
- Trust & Adoption – Convincing medical professionals to trust AI-driven predictions.

5. Expected Outcomes & Benefits

- Early Detection – Helps doctors intervene before severe liver damage occurs.
- Healthcare Efficiency – Reduces burden on hospitals by prioritizing high-risk cases.
- Patient Awareness – Educates individuals about cirrhosis risk and lifestyle modifications.

REQUIREMENT ANALYSIS

Customer Journey Map



Solution Requirement

To ensure the successful implementation of the predictive model, the following functional and nonfunctional requirements must be met:

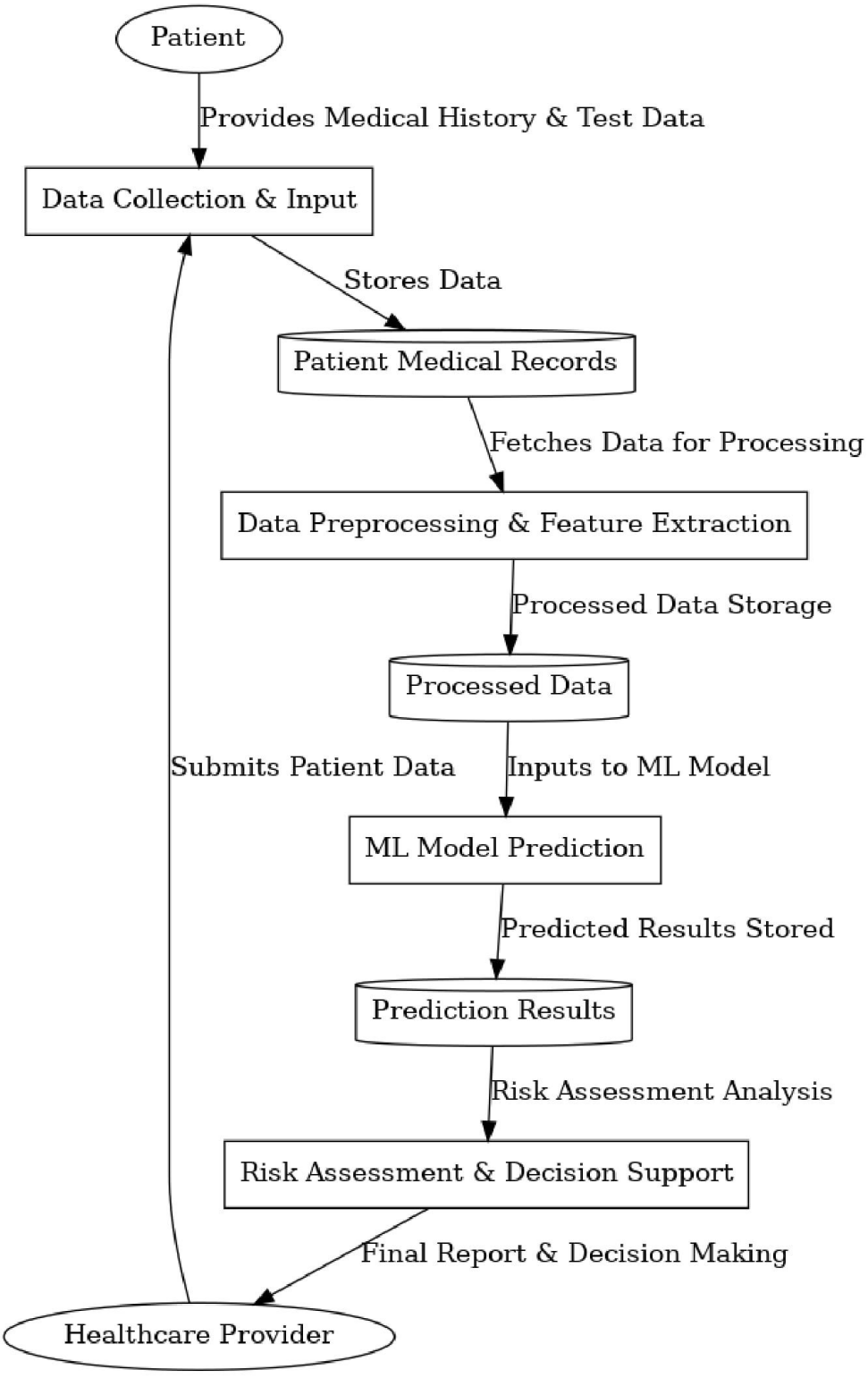
Functional Requirements:

- Integration with Electronic Health Records (EHR) systems.
- Data preprocessing to clean and normalize medical records.
- Machine learning model capable of classifying risk levels based on patient data.
- Real-time or batch processing of new patient data.
- User-friendly interface for healthcare professionals to interpret results.
- Automated alerts for high-risk patients.

Non-Functional Requirements:

- High accuracy and reliability of predictions.
- Data security and compliance with healthcare regulations (e.g., HIPAA, GDPR).
- Scalability to handle increasing patient records over time.
- Low-latency processing for quick decision-making.
- Regular updates and retraining of the model with new medical data.

Data Flow Diagram



Technology Stack

To develop and deploy the predictive liver cirrhosis model efficiently, the following technology stack will be used:

Programming Languages

- Python (for machine learning, data preprocessing, and backend) Data Handling & Storage
- Data Processing Libraries: Pandas, NumPy, SciPy
- Data Visualization: Matplotlib, Seaborn Machine Learning & AI
- ML Frameworks: Scikit-learn, XGBoost, TensorFlow/PyTorch
- Models Tested: Logistic Regression, DecisionTree, RandomForest, XGBoost, Support Vector Classifier, KNeighboursClassifier, Gaussian Naïve Bayes
- Model Deployment: Flask (as API layer)
- Saving Model: Pickle

PROJECT DESIGN

Problem Solution Fit

Liver cirrhosis is a progressive disease that is often detected at an advanced stage, leading to complications and high healthcare costs. Traditional diagnostic methods rely on symptomatic evaluation and expensive imaging techniques, which may delay early intervention. By leveraging machine learning, this project offers a data-driven approach to predicting cirrhosis risk at an early stage, enabling timely medical intervention and reducing the burden on healthcare systems.

Proposed Solution:

The proposed solution is a machine learning-based predictive model that assesses cirrhosis risk using patient data.

The key components include:

- Data Acquisition: Collecting structured and unstructured patient records, including lab tests, imaging results, and medical history.
- Feature Engineering: Identifying critical biomarkers and patient attributes that contribute to cirrhosis risk.
- Model Training: Implementing supervised learning techniques to classify patients based on risk levels.
- Deployment & Integration: Embedding the model within healthcare IT systems for real-time predictions.
- Continuous Monitoring: Updating the model periodically with new data to improve predictive accuracy.

Solution Architecture

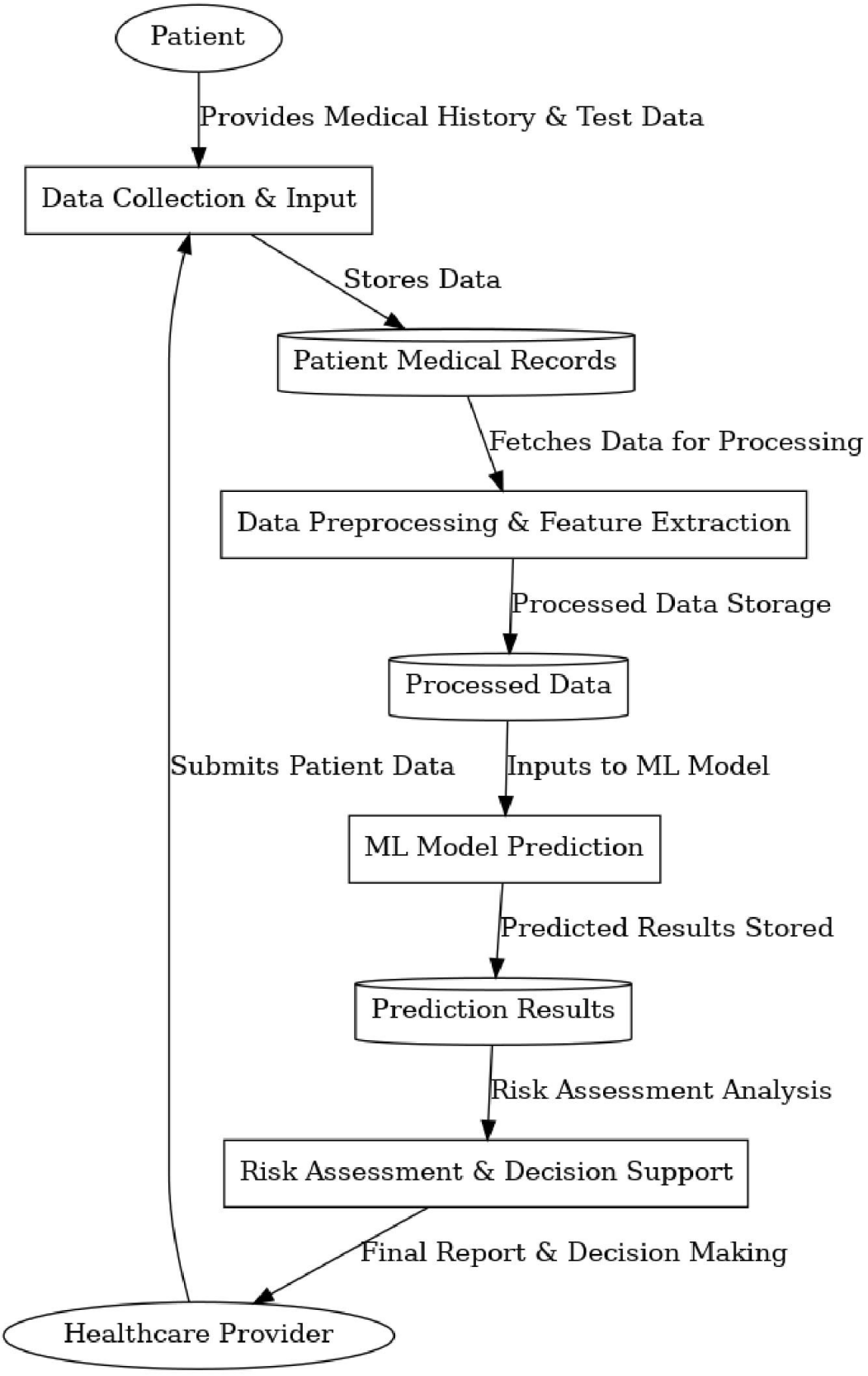
The architecture of the predictive system comprises multiple layers to ensure seamless data processing and model inference:

1. Data Ingestion Layer – Aggregates data from EHR systems, lab reports, and patient history.

2. Data Processing Layer – Cleans and transforms raw data into a structured format for machine learning.

3. Model Training & Inference Layer – Trains ML models using historical patient data and generates real-time predictions.

4. Storage & Monitoring Layer – Stores patient data securely while tracking model performance for continuous improvement.


Patient
Provides Medical History & Test Data
Data Collection & Input
Stores Data
Patient Medical Records
Fetches Data for Processing
Data Preprocessing & Feature Extraction
Processed Data Storage
Processed Data
Submits Patient Data
Inputs to ML Model
ML Model Prediction
Predicted Results Stored
Prediction Results
Risk Assessment Analysis
Risk Assessment & Decision Support
Final Report & Decision Making
Healthcare Provider

PROJECT PLANNING & SCHEDULING

Project Planning

Project Planning Map

1. Problem Definition & Research

- Identify key challenges in liver cirrhosis diagnosis.
- Analyse existing diagnostic methods and their limitations.
- Define project objectives and success criteria.

2. Data Collection & Preprocessing

- Gather structured & unstructured medical data (EHR, lab tests, imaging).
- Handle missing values, outliers, and normalize data.
- Feature selection: Identify key biomarkers and relevant attributes. 3. Model Development
- Choose ML algorithms (Random Forest, XGBoost, Neural Networks, etc.).
- Train and validate models using historical patient data.
- Optimize hyperparameters for best performance. 4. Model Evaluation & Validation
- Measure performance using metrics: Accuracy, Precision, Recall, F1-score, AUC-ROC.
- Compare multiple models to select the best one.
- Interpret feature importance for explainability.

5. Continuous Monitoring & Updates

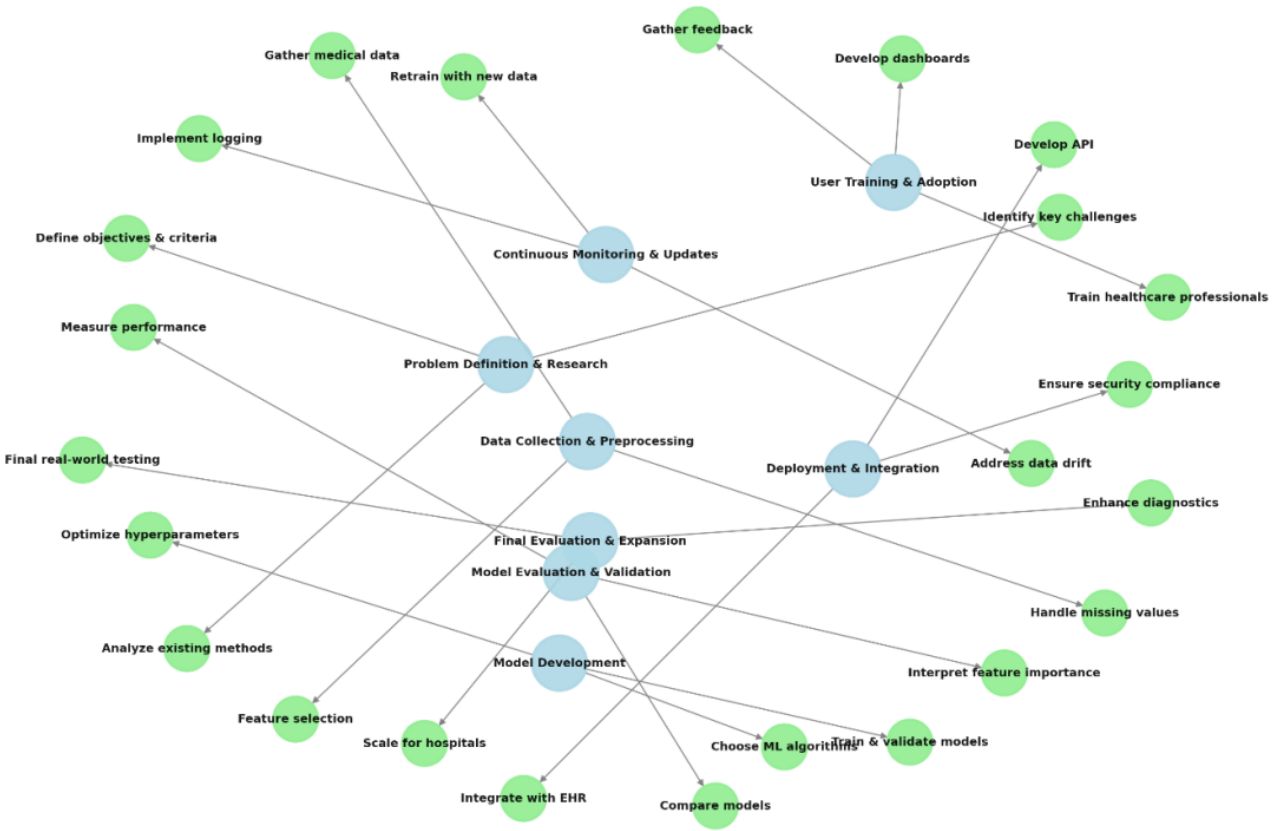
- Implement logging & performance tracking.
- Periodically retrain model with new patient data.
- Address data drift and improve prediction accuracy.

7. User Training & Adoption

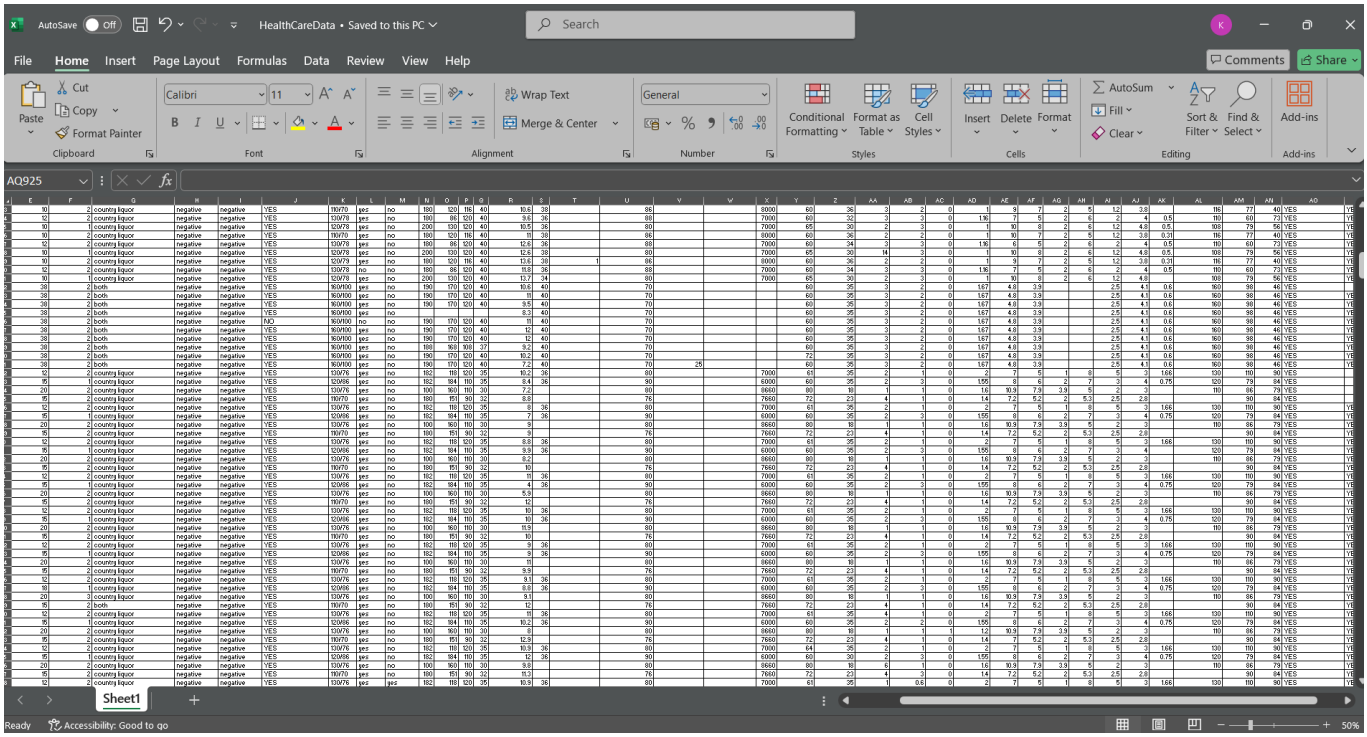
- Provide training to healthcare professionals.
- Develop user-friendly dashboards for easy interpretation.
- Gather feedback for iterative improvements.

8. Final Evaluation & Expansion

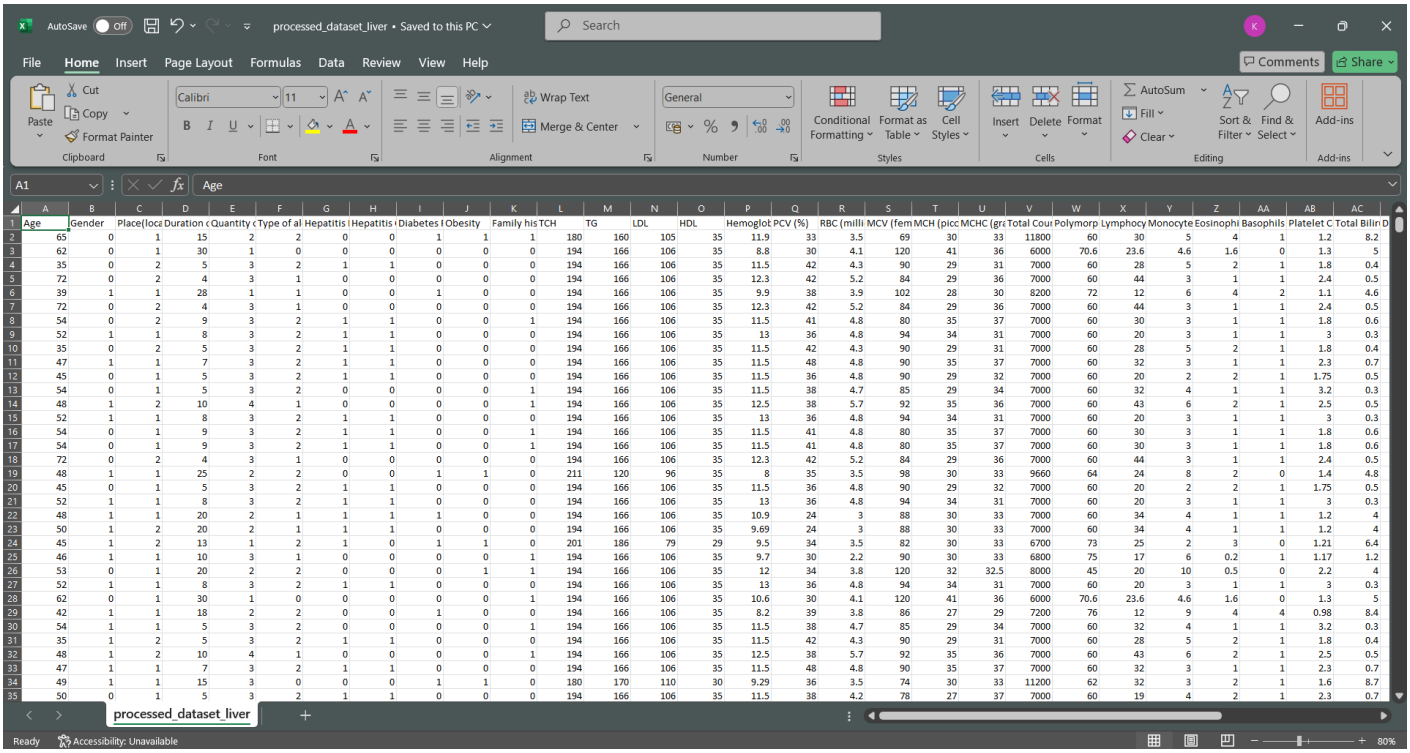
- Conduct final testing with real-world patient data.
- Scale the system for broader adoption in hospitals.
- Explore additional predictive features for enhanced diagnostics.



FUNCTIONAL PERFORMING &TESTING



HealthCareData.xls



processed_dataset_liver.csv

In [1]:
```
import pandas as pd
import numpy as np
import openpyxl

from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, MinMaxScaler, LabelEncoder, OneHotEncoder
from sklearn.utils import resample
```

In [2]:
```
df= pd.read_excel('Dataset/HealthCareData.xlsx', engine='openpyxl')
```

In [3]:
```
pd.set_option('display.max_columns', None)
```

In [4]:
```
df.head()
```

Out[4]:

	S.NO	Age	Gender	Place(location where the patient lives)	Duration of alcohol consumption(years)	Quantity of alcohol consumption (quarters/day)	Type of alcohol consumed	Hepatitis B infection	Hepatitis C infection	Diabetes Result	Blood pressure (mmhg)	Obesity
0	1	55	male	rural	12	2	branded liquor	negative	negative	YES	138/90	yes
1	2	55	male	rural	12	2	branded liquor	negative	negative	YES	138/90	yes
3	4	55	male	rural	12	2	branded liquor	negative	negative	NO	138/90	no
4	5	55	female	rural	12	2	branded liquor	negative	negative	YES	138/90	no

In [5]:
```
df.shape
```

Out[5]:
```
(950, 42)
```

In [6]:
```
df.describe()
```

Out[6]:

	S.NO	Age	Duration of alcohol consumption(years)	Quantity of alcohol consumption (quarters/day)	TCH	HDL	Hemoglobin (g/dl)	PCV (%)	RBC (million cells/microliter)	(f
count	950.000000	950.000000	950.000000	950.000000	591.000000	582.000000	950.000000	920.000000	398.000000	
mean	475.500000	50.632632	20.606316	5.158947	197.544839	35.486254	10.263979	33.810000	3.390704	
std	274.385677	8.808272	7.980664	22.908785	26.694968	7.982057	1.942300	5.751592	0.937089	
min	1.000000	32.000000	4.000000	1.000000	100.000000	25.000000	4.000000	12.000000	1.000000	
25%	238.250000	44.000000	15.000000	2.000000	180.000000	30.000000	9.000000	30.000000	2.825000	
50%	475.500000	50.000000	20.000000	2.000000	194.000000	35.000000	10.000000	35.000000	3.500000	
75%	712.750000	57.000000	26.000000	3.000000	210.000000	38.000000	11.500000	38.000000	4.000000	

data_preprocessing.ipynb

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [3]:
```
csv_file= "processed_dataset_liver.csv"
df = pd.read_csv(csv_file)
display(df.info())
display(df.head())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1860 entries, 0 to 1859
Data columns (total 42 columns):
 #   Column                                                                Non-Null Count  Dtype
---  ------                                                                --------------  -----
 0   Age                                                                   1860 non-null   int64
 1   Gender                                                                1860 non-null   int64
 2   Place(location where the patient lives)                               1860 non-null   int64
 3   Duration of alcohol consumption(years)                                1860 non-null   int64
 4   Quantity of alcohol consumption (quarters/day)                        1860 non-null   float64
 5   Type of alcohol consumed                                              1860 non-null   int64
 6   Hepatitis B infection                                                 1860 non-null   int64
 7   Hepatitis C infection                                                 1860 non-null   int64
 8   Diabetes Result                                                       1860 non-null   int64
 9   Obesity                                                               1860 non-null   int64
 10  Family history of cirrhosis/ hereditary                               1860 non-null   int64
 11  TCH                                                                   1860 non-null   float64
 12  TG                                                                    1860 non-null   float64
 13  LDL                                                                   1860 non-null   float64
 14  HDL                                                                   1860 non-null   float64
 15  Hemoglobin (g/dl)                                                     1860 non-null   float64
 16  PCV (%)                                                               1860 non-null   float64
 17  RBC (million cells/microliter)                                        1860 non-null   float64
 18  MCV (femtoliters/cell)                                                1860 non-null   float64
 19  MCH (picograms/cell)                                                  1860 non-null   float64
 20  MCHC (grams/deciliter)                                                1860 non-null   float64
 21  Total Count                                                           1860 non-null   float64
 22  Polymorphs (%)                                                        1860 non-null   float64
 23  Lymphocytes (%)                                                       1860 non-null   float64
 24  Monocytes (%)                                                         1860 non-null   float64
 25  Eosinophils (%)                                                       1860 non-null   float64
 26  Basophils (%)                                                         1860 non-null   float64
 27  Platelet Count (lakhs/mm)                                             1860 non-null   float64
 28  Total Bilirubin (mg/dl)                                               1860 non-null   float64
 29  Direct (mg/dl)                                                        1860 non-null   float64
 30  Indirect (mg/dl)                                                      1860 non-null   float64
 31  Total Protein (g/dl)                                                  1860 non-null   float64
 32  Albumin (g/dl)                                                        1860 non-null   float64
 33  Globulin (g/dl)                                                       1860 non-null   float64
 34  A/G Ratio                                                             1860 non-null   float64
 35  AL.Phosphatase (U/L)                                                  1860 non-null   float64
 36  SGOT/AST (U/L)                                                        1860 non-null   int64
 37  SGPT/ALT (U/L)                                                        1860 non-null   int64
 38  USG Abdomen (diffuse liver or not)                                    1860 non-null   int64
 39  Predicted Value(Out Come-Patient suffering from liver cirrosis or not)  1860 non-null   int64
 40  Systolic BP                                                           1860 non-null   int64
 41  Diastolic BP                                                          1860 non-null   int64
dtypes: float64(26), int64(16)
memory usage: 610.4 KB
None
```

EDA.ipynb

In [1]:

```
import pandas as pd
import numpy as np

import seaborn as sns
import matplotlib.pyplot as plt

from sklearn.preprocessing import StandardScaler

from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier, plot_tree
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier
from sklearn.svm import SVC
from sklearn.neighbors import KNeighborsClassifier
import xgboost as xgb
from sklearn.naive_bayes import GaussianNB

from sklearn.model_selection import train_test_split, GridSearchCV, RandomizedSearchCV, cross_val_score, StratifiedKFold
from sklearn.metrics import classification_report, accuracy_score, precision_score, recall_score, f1_score, roc_auc_score
from sklearn.feature_selection import RFECV
import optuna
from optuna import create_study

import pickle
```

```
b:\Git hub\Revolutionizing_Liver_Care_-_Predicting_Liver_Cirrhosis_Using_Advanced_Machine_Learning_Techniques\.env\Lib\site
-packages\tqdm\auto.py:21: TqdmWarning: IProgress not found. Please update jupyter and ipywidgets. See https://ipywidgets.r
eadthedocs.io/en/stable/user_install.html
  from .autonotebook import tqdm as notebook_tqdm
  from .autonotebook import tqdm as notebook_tqdm
```

In [2]:

```
pd.set_option('display.max_columns', None)
```

In [3]:

```
df=pd.read_csv('Dataset/processed_dataset.csv')
df.head()
```

Out[3]:

	Age	Gender	Place(location where the patient lives)	Duration of alcohol consumption(years)	Quantity of alcohol consumption (quarters/day)	Type of alcohol consumed	Hepatitis B infection	Hepatitis C infection	Diabetes Result	Obesity	Family history of cirrhosis/ hereditary	
0	65	0	1	15	2.0	2	0	0	1	1	1	1
1	62	0	1	30	1.0	0	0	0	0	0	0	1
2	35	0	2	5	3.0	2	1	1	0	0	0	1
3	72	0	2	4	3.0	1	0	0	0	0	0	1
4	39	1	1	28	1.0	1	0	0	1	0	0	1

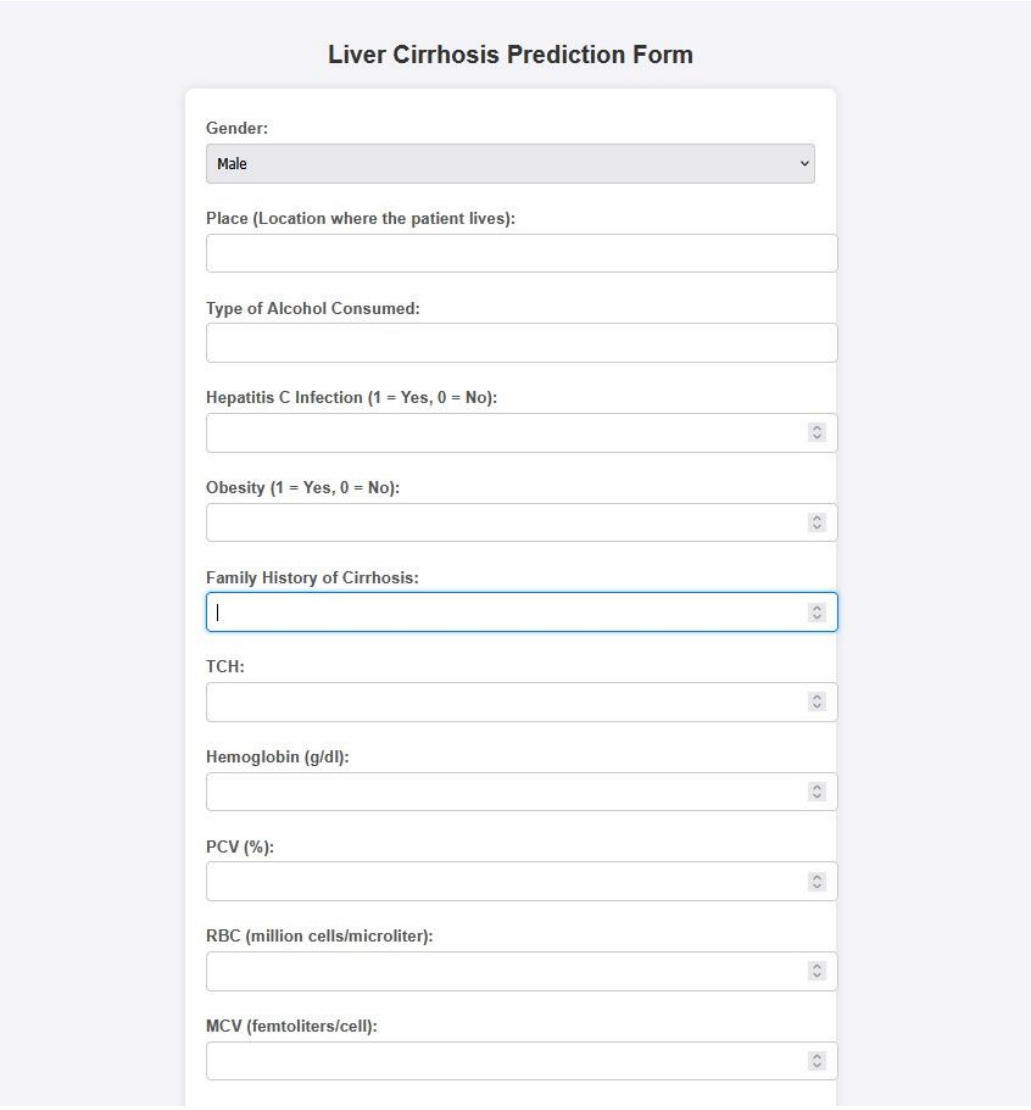
In [4]:

```
df.drop_duplicates(inplace=True)
df.duplicated().sum()
```

model_building.ipynb

```
pandas
numpy
seaborn
matplotlib
scikit-learn
xgboost
optuna
pickle5  # Only required for Python versions < 3.8
```

requirements.txt



flask web integration UI

MCH (picograms/cell):

MCHC (grams/deciliter):

Total Count:

Polymorphs (%):

Lymphocytes (%):

Monocytes (%):

Eosinophils (%):

Basophils (%):

Platelet Count (lakhs/mm):

Indirect Bilirubin (mg/dl):

Total Protein (g/dl):

Albumin (g/dl):

Globulin (g/dl):

Liver Stress Score:

Metabolic Syndrome Indicator:

Family History of Diabetes:

Obesity LDL:

Obesity TG:

Age Category (Middle-Aged):

Age Category (Young):

Alcohol Category (Light):

Alcohol Category (Moderate):

Alcohol Category (Heavy):

BP Category (Prehypertension):

BP Category (Hypertension):

Predict

ADVANTAGES & DISADVANTAGES

Advantages & Disadvantages

Advantages

1. **Early Detection & Prevention** – The model enables early identification of liver cirrhosis, allowing timely medical intervention to slow disease progression and improve patient outcomes.
2. **Non-Invasive Diagnosis** – Unlike biopsies and certain imaging techniques, this AI-driven approach relies on existing medical and lifestyle data, reducing the need for invasive procedures.
3. **Data-Driven Decision Making** – The predictive model provides objective insights based on vast amounts of clinical data, supporting healthcare professionals in making more informed treatment decisions.

Disadvantages

1. **Data Availability & Quality Issues** – The model's accuracy depends on high-quality, diverse clinical data. Incomplete or biased datasets could lead to unreliable predictions.
2. **Interpretability & Trust Issues** – Many AI models operate as "black boxes," making it difficult for doctors to understand how predictions are generated, which may affect trust in the system.
3. **Regulatory & Ethical Concerns** – AI-driven medical tools must comply with strict healthcare regulations (e.g., HIPAA, GDPR) regarding data privacy and patient safety, posing implementation challenges.

CONCLUSION

Conclusion

This project aims to revolutionize liver disease management by leveraging advanced machine learning techniques to predict liver cirrhosis risk accurately and efficiently. By analyzing diverse patient data—including medical history, laboratory results, imaging scans, and lifestyle factors—the predictive model provides a non-invasive, data-driven approach to early detection and disease progression monitoring.

The integration of this model into a user-friendly web application enhances accessibility for healthcare professionals, enabling real-time risk assessment and informed decision-making. This can lead to earlier interventions, personalized treatment plans, and improved patient outcomes while reducing the reliance on invasive and expensive diagnostic methods.

Despite challenges such as data quality, model interpretability, and regulatory considerations, this project represents a significant step toward AI-driven predictive healthcare. By bridging the gap between technology and medicine, it offers a scalable and cost-effective solution that has the potential to positively impact patient care and healthcare systems worldwide.

Moving forward, continuous refinement of the model, integration with electronic health records, and collaboration with medical professionals will be crucial to ensuring accuracy, reliability, and ethical deployment in real-world clinical settings.

FUTURE SCOPE

Future Scope

 Enhancing Model Accuracy with More Data – Expanding the dataset with more diverse patient records from different demographics, hospitals, and geographic regions can improve the model's accuracy and generalizability.

 Integration with Electronic Health Records (EHRs) – Connecting the predictive model with hospital EHR systems would enable seamless data input and real-time risk assessments, improving workflow efficiency for healthcare professionals.

 AI Explainability & Transparency – Developing explainable AI (XAI) techniques to make model predictions more interpretable for doctors and patients will help increase trust and adoption in clinical settings.

 Real-Time Monitoring & Predictive Analytics – Implementing continuous monitoring of high-risk patients using wearables and IoT-based health devices could provide real-time risk predictions and early alerts for disease progression.

 Multi-Disease Prediction Capability – Expanding the model to predict other liver diseases (e.g., hepatitis, fatty liver disease, or liver cancer) could make it a more comprehensive diagnostic tool.

 Mobile & Cloud-Based Deployment – Developing a mobile-friendly or cloud-based version of the application would improve accessibility, allowing remote diagnostics and telemedicine integration for underserved areas.

 Personalized Treatment Recommendations – By incorporating AI-driven treatment suggestions based on patient-specific risk factors,

the system could support precision medicine approaches in hepatology.

 Regulatory Approval & Clinical Trials – Conducting clinical trials and obtaining necessary healthcare regulatory approvals (e.g., FDA, CE certification) will be essential for real-world deployment in hospitals and medical institutions.

APPENDIX

Appendix

Dataset Link: Kaggle

Github: Project Repository Link

Project Demo Link: Google Drive Link