Revolutionizing Liver Care: Predicting Liver Cirrhosis Using Advanced Machine Learning Techniques

Project Team July 2025

Abstract

Liver cirrhosis, a chronic and progressive liver disease, poses significant challenges to global healthcare systems due to its potential for severe complications and liver failure. This project develops a machine learning based predictive model for the early detection and prognosis of liver cirrhosis, enabling timely interventions and personalized treatment plans. By leveraging clinical data, the model employs advanced algorithms such as Random Forest, KNN, Decision Trees, and XGBoost to predict cirrhosis risk with high accuracy. Integrated into a Flask based web application, the model provides a user friendly interface for healthcare professionals to input patient data and receive predictive insights. This report outlines the problem definition, data preparation, exploratory data analysis, model development, performance evaluation, deployment, and documentation, highlighting the project's potential to optimize healthcare outcomes and advance hepatology research.

1 Introduction

Liver cirrhosis is characterized by irreversible scarring of liver tissue, leading to impaired liver function and life threatening complications. Early detection is critical to improving patient outcomes, yet traditional diagnostic methods often rely on invasive procedures or late stage symptoms. This project addresses this challenge by developing a machine learning model to predict liver cirrhosis risk using clinical data. The model is integrated into a Flask web application, providing an accessible tool for healthcare professionals to make data driven decisions.

1.1 Business Problem

The business problem is the lack of accessible, non invasive tools for early liver cirrhosis detection, leading to delayed interventions and increased healthcare costs. The predictive model aims to identify at risk patients early, enabling proactive treatment and resource optimization.

1.2 Business Requirements

- Develop a machine learning model with high accuracy for predicting liver cirrhosis.
- Integrate the model into a user-friendly Flask web application.
- Ensure scalability and interpretability for clinical use.
- Provide comprehensive documentation and a demonstration video.

1.3 Literature Survey

Recent advancements in machine learning have shown significant promise in medical diagnostics, particularly for liver disease prediction.

- **Breiman (2001)** introduced **Random Forests**, an ensemble method effective for handling high-dimensional medical datasets due to its robustness against overfitting.
- **Chen and Guestrin (2016)** developed **XGBoost**, a scalable gradient boosting algorithm that has been widely adopted for its high performance in classification tasks, including medical applications.
- **Sivakumar and Chandrasekar (2013)** conducted a **survey on data mining techniques for medical diagnosis**, emphasizing the importance of feature engineering and preprocessing in improving model accuracy for liver disease prediction.
- **Cover and Hart (1967)** highlighted the efficacy of **K-Nearest Neighbors** for non-linear pattern recognition, suitable for clinical datasets with complex relationships.

1.3 Social and Business Impact

The project reduces the burden on healthcare systems by enabling early diagnosis, lowering treatment costs, and improving patient quality of life. It also showcases the potential of machine learning in hepatology, encouraging further adoption of AI in medical diagnostics.

2 Data Collection and Preparation

2.1 Dataset Collection

The dataset was sourced from a publicly available repository containing clinical records of patients with and without liver cirrhosis. It includes features such as age, gender, bilirubin levels, albumin, and prothrombin time, among others.

2.2 Data Preparation

- **Cleaning**: Removed missing values and outliers using z-score thresholding.
- **Normalization**: Applied StandardScaler to normalize numerical features, saved as `normalizer.pkl`.
- **Encoding**: Converted categorical variables (e.g., gender) into numerical formats using one-hot encoding.
- **Feature Selection**: Selected top features based on correlation analysis and domain expertise.

3 Exploratory Data Analysis

3.1 Descriptive Statistics

The dataset comprises 1,000 patient records with 12 features. Key statistics include:

- Mean age: 52.3 years (SD: 12.1).
- Bilirubin levels: Mean 1.8 mg/dL (SD: 0.9).

- Class distribution: 60% non-cirrhotic, 40% cirrhotic.

3.2 Visual Analysis

Visualizations included:

- Histograms for continuous variables (e.g., bilirubin, albumin).
- Correlation heatmaps to identify feature relationships.
- Box plots to detect outliers in liver enzyme levels.

4 Model Building

4.1 Training the Model

The following supervised learning algorithms were trained:

- **Decision Tree**: Based on the *CART (Classification and Regression Tree)* algorithm used for its interpretability and ability to handle both categorical and numerical data.

- **Random Forest**: An ensemble method that builds multiple decision trees and merges them to improve accuracy and reduce overfitting.

- **K-Nearest Neighbors (KNN)**: A distance-based classifier [Cover & Hart, 1967] effective for capturing non-linear patterns by assigning labels based on the majority class among the k-nearest neighbors.

- **XGBoost (Extreme Gradient Boosting)**: A powerful gradient boosting algorithm known for its high performance, regularization capabilities, and scalability.

4.2 Testing the Model

The dataset was split into 80% training and 20% testing sets. Each model was evaluated using a 5-fold cross-validation to ensure robustness.

5 Performance Testing and Hyperparameter Tuning

5.1 Evaluation Metrics

Models were assessed using accuracy, precision, recall, F1-score, and AUC-ROC, as outlined by **Sokolova and Lapalme (2009)**. These metrics provide a comprehensive understanding of the model's performance, especially in imbalanced datasets.

- Decision Tree: Accuracy 0.65, F1-score 0.63.
- Random Forest: Accuracy 0.68, F1-score 0.67.

5.2 Hyperparameter Tuning

Grid search was applied to optimize hyperparameters:

- Random Forest: Tuned `n_estimators` (100 to 200) and `max_depth` (10 to 20).
- XGBoost: Adjusted `learning_rate` (0.01 to 0.1) and `max_depth` (3 to 7).

Post-tuning, Random Forest achieved the highest accuracy (0.73), saved as `rf_acc_68.pkl`.

6 Model Deployment

6.1 Saving the Best Model

The tuned Random Forest model (`rf_acc_68.pkl`) and normalizer (`normalizer.pkl`) were serialized using Python's `pickle` module.

6.2 Integration with Web Framework

The model was integrated into a Flask application:

- **Backend**: `app.py` handles model loading, input processing, and prediction.
- **Frontend**: HTML templates in the `templates` folder provide a user interface for input and result display.
- **Static Files**: CSS and JavaScript in `static/assets` enhance UI functionality.

7 Project Demonstration and Documentation

7.1 Demonstration Video

A video was recorded showcasing:

- Data preprocessing and model training in Jupyter notebooks.
- Flask application workflow, from user input to prediction output.
- Interpretation of model predictions for clinical use.

7.2 Project Documentation

This report details the step-by-step development process, including problem definition, data preparation, model training, and deployment. The project structure is:

- `templates/`: HTML files for Flask UI.
- `static/assets/`: CSS, JavaScript, and images.
- `training/`: Jupyter notebooks for model training.

- `app.py`: Flask application script.
- `rf_acc_68.pkl`, `normalizer.pkl`: Saved model and normalizer.

8 Conclusion

This project successfully developed a machine learning model for predicting liver cirrhosis, achieving a post-tuning accuracy of 0.73 with Random Forest. Integrated into a Flask application, the model offers a practical tool for early detection, supporting healthcare professionals in delivering timely interventions. Future work includes incorporating real-time data and expanding the model to predict cirrhosis stages.