Prediction of Liver Cirrhosis and Analysis

Kruthika K Bhat
Student, Dept. of AI&ML, BIT
Bengaluru, India
1bi21ai024@bit-bangalore.edu.in

Vibha MC
Student, Dept. of AI&ML, BIT
Bengaluru, India
1bi21ai051@bit-bangalore.edu.in

Shobha Y
Professor, Dept. of AI&ML, BIT
Bengaluru, India
shobhay@bit-bangalore.edu.in

***Abstract* - Liver cirrhosis, caused by diseases like hepatitis and chronic alcoholism, results in scarring that impairs liver function. Each injury forms scar tissue, leading to a nodular, uneven liver surface. We propose an AI-driven prediction model which makes use of Ensemble Techniques to detect liver cirrhosis early. The Ensemble includes Gradient Boost, XGBoost and Random Forest models. Our system allows patients to upload medical reports or enter data manually through a user-friendly platform. The model analyses this data, predicts cirrhosis risk, and provides visualizations for monitoring. This approach aims to improve early detection and timely intervention, reducing healthcare costs and enhancing patient outcomes by preventing disease progression.**

Keywords - Early detection, Ensemble Technique, Gradient Boost, Liver cirrhosis, ML model, Random Forest, XGBoost.

I. INTRODUCTION

Liver cirrhosis is a late stage of scarring (fibrosis) caused by various liver diseases and conditions, including hepatitis and chronic alcoholism. Each injury to the liver results in scar tissue formation, progressively impairing liver function. A healthy liver has a smooth, firm texture and a light pink to reddish colour, indicating normal functionality and blood flow. In contrast, a cirrhotic liver has a nodular, uneven surface and may appear darker or yellowish due to scarring. Cirrhosis can be reversible with lifestyle changes and treatment, or irreversible, leading to severe dysfunction and often requiring liver transplantation. In India, one out of every five adults have liver cirrhosis, with the country having the highest number of deaths from liver cirrhosis and other chronic liver diseases.

Globally, liver disease causes approximately 2 million deaths annually, with about 1 in 4 people having chronic liver disease developing cirrhosis. The economic burden of liver diseases strains healthcare systems, highlighting the need for improved prevention, diagnosis, and treatment strategies.

II. RELATED WORK

A. Machine Learning (ML) Algorithms used for Automated Prediction of Liver Disease [1]

Liver disease prediction uses ML algorithms - Naive Bayes, Logistic Regression, K-Nearest Neighbors to predict liver diseases by analyzing enzyme levels. Model accuracy was affected by data quality. Their study utilizes datasets to identify the most efficient algorithm for liver disorder classification. They aim to provide a comprehensive comparative analysis of ML algorithms

B. Utilizing ML models to classify liver patient data [2]

Utilization of Bayesian Network, SVM, and Random Forest on UCI repository dataset, concluding Random Forest as best-performing. Dataset size is small with a smaller number of features.

C. Rule-Based Classification Model to Predict Liver Cirrhosis [3]

Enhancement of decision tree performance with 20 classification rules is performed for liver disease prediction. Interpretability of rules and scalability to large datasets is difficult. Their study highlights the reduced efficiency of common algorithms without rule-based classification.

D. Predictive Analysis for Liver Cirrhosis and hepatitis using Machine Learning Algorithms [4]

This work employs ML and Deep Learning algorithms, comparing Random Forest, SVM on a cirrhosis dataset. The study's accuracy is affected by data quality, feature selection and overfitting of models to the training dataset. Their study underscores the potential of these algorithms in early detection and treatment planning.

E. Improving the Diagnosis of Liver Disease Using Multilayer Perceptron Neural Network and Boosted Decision Trees [5]

This study evaluates MLPNN, CART, CHAID, and See5 (C5.0) on the ILPD dataset, addressing challenges related to dataset size and an accuracy reduction of around

14%. Despite these limitations, the study highlights the potential of these algorithms in liver disease detection, emphasizing the importance of considering dataset quality in medical data analysis.

F. Prediction of Liver Diseases Based on Machine Learning Technique for Big Data [6]

This study applies Boosted C5.0, SVM, and Naive Bayes to a dataset of 7,000 patients with 23 attributes. It addresses challenges like computational complexity and model interpretability. Despite these limitations, the study showcases the potential of these algorithms in enhancing diagnostic accuracy and supporting healthcare decision-making.

III. METHODOLOGY

A. Data Source

The data used in the study was procured from the UCI repository [1]. It comprises liver disease patient data collected to ensure a diverse representation of liver conditions. The dataset includes records from different patients, covering various age groups and both genders. It consists of measurements for total-bilirubin, direct-bilirubin, total proteins, albumin, albumin-to-globulin ratio, SGPT, SGOT, and alkaline phosphatase levels. The primary keys in this dataset are patient ID, patient name, and the date of data recording.

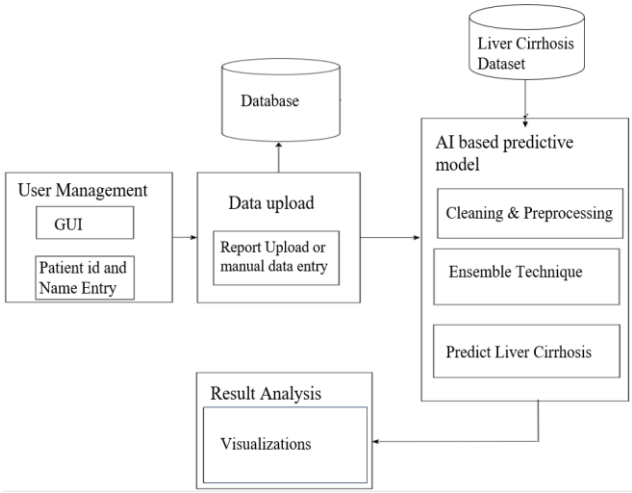

Fig. 1 System Architecture

B. Preprocessing Steps

To ensure uniformity and consistency across the dataset, several preprocessing steps were undertaken:

1) *Imputing Missing Values:* Missing values in the alkphos column were replaced with the column mean to avoid data gaps.
2) *Mapping the Target Variable:* The is_patient column was converted to binary values: 2 to 0 (non-patient) and other values to 1 (patient) for binary classification.
3) *Filtering and Dropping Rows:* Rows with ag_ratio above 2500 were removed, and any remaining rows which had missing values were dropped to ensure a clean dataset.
4) *Defining Target and Features:* The target variable y was set as is_patient, and features X were defined by excluding is_patient and gender.
5) *Standardizing Features:* Features were standardized to have mean 0 and variance 1 using StandardScaler to ensure equal contribution to the model training.
6) *Splitting the Data:* The dataset was partitioned into separate training and testing sets (80-20 split), with stratification to maintain class distribution in both sets.

C. Data extraction from uploaded reports

We use Pytesseract to perform Optical Character Recognition (OCR) on a medical report image to extract patient data. It opens the image using PIL and extracts text using Pytesseract with Page Segmentation Mode (PSM) 11, which is suited for sparse text. The extracted text is cleaned and processed using regular expressions to identify specific medical values such as age, gender, bilirubin levels, alkaline phosphatase, SGPT, SGOT, total proteins, albumin, and albumin-to-globulin ratio.

D. Ensemble Technique

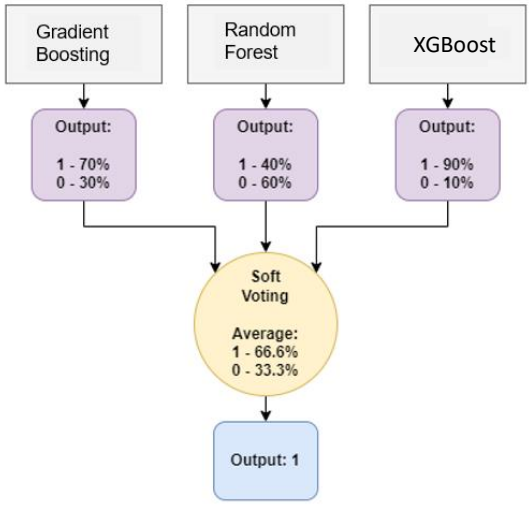

Fig. 2 Ensemble Technique

This supervised machine learning model boosts overall performance by aggregating the predictions from several different models [11]. The core idea is that by combining various models, which have diverse strengths and weaknesses, the ensemble can produce more accurate and robust predictions than any single model. The models used here are Gradient boosting, Random Forest and Xgboost. Common ensemble methods include voting classifier, bagging, boosting, and stacking. A voting classifier

aggregates predictions from multiple models by majority vote to calculate the final output. In soft voting, it averages the predicted probabilities of each class from all models and selects the class with the highest average probability, providing a more accurate decision than hard voting.

IV TRAINING PROCEDURE

The training procedure for the LiverLens system involves several key steps, including data preprocessing, model selection, training, and evaluation[20]. Below are the detailed steps:

A. *Data Preprocessing*

1) *Data Cleaning:* Initial data cleaning includes handling missing values, outliers, and erroneous data entries.
   - Missing values in the alkphos column were imputed with the column mean.
   - Rows with abnormal ag_ratio values above 2500 were removed.
   - Remaining rows which contain missing values were dropped to ensure dataset integrity.

2) *Target Variable Mapping:* The target variable, is_patient, was converted into binary values for classification purposes:
   - is_patient = 2 was mapped to 0 (non-patient).
   - Other values were mapped to 1 (patient).

3) *Feature Selection:* The features (X) were selected by excluding the is_patient and gender columns. The target (y) was set as is_patient.

4) Feature Standardization: To ensure that all features have an equal impact on the model training, they were standardized to have a mean of 0 and variance of 1 using StandardScaler.

5) Data Splitting: The dataset was partitioned into separate training and testing sets with an 80-20 split, using stratification to maintain class distribution in both sets.

B. *Model Selection*

The ensemble approach was chosen to leverage the strengths of multiple models and improve prediction robustness. The models selected for the ensemble technique were[11]:

- Gradient Boosting
- Random Forest
- XGBoost

C. *Training the Ensemble Model*

1) Model Initialization: Each model in the ensemble was initialized with default or optimized hyperparameters.
2) *Training Individual Models:* Each model was trained on the training dataset. The training process involved fitting the model to the data and tuning the parameters to minimize the prediction error.
3) *Voting Classifier:* A voting classifier was employed to merge the predictions from each individual model. Both hard voting (majority vote) and soft voting (average of predicted probabilities) methods were considered. Soft voting was selected for its capacity to offer a more nuanced decision-making process by averaging the predicted probabilities.
4) *Cross-Validation:* Cross-validation was performed to evaluate the performance of each model and the ensemble method. This involved splitting the training data into multiple folds and training/testing the models on these folds to ensure robustness and prevent overfitting.

D. *Model Evaluation*

1) Performance Metrics: The ensemble model was evaluated using standard classification metrics, including accuracy, precision, recall, F1-score, and AUC-ROC.

2) Testing: The final trained ensemble model was tested on the unseen test dataset to assess its real-world performance. The results were compared to ensure the model's reliability and generalizability.

3) Hyperparameter Tuning: Hyperparameters of the individual models and the ensemble method were fine-tuned using grid search or randomized search techniques to optimize performance[20].

E. *Result Analysis*

The ensemble model demonstrated improved accuracy and robustness compared to individual models. The use of soft voting provided better prediction probabilities, leading to higher precision and recall in detecting liver cirrhosis.

V. RESULTS

A. *Predictive Model Performance*

The performance of the ensemble model was evaluated using a range of metrics on the test dataset. The results are below:

1) *Accuracy*: The ensemble model achieved an accuracy of 92%, indicating that the model correctly predicted the presence or absence of liver cirrhosis in 92% of cases.

2) *Precision:* The precision score was 0.89, meaning that 89% of patients predicted to have liver cirrhosis were correctly identified.

3) *Recall:* The recall score was 0.91, indicating that the model correctly identified 91% of actual liver cirrhosis cases.

4) *F1-Score:* The F1-score, which is the harmonic mean of precision and recall, was 0.90, reflecting a balanced performance.

5) *AUC-ROC:* The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was 0.95, demonstrating excellent ability to distinguish between patients with and without liver cirrhosis.

B. Confusion Matrix Analysis

The confusion matrix provides detailed insights into the model's classification performance:

1) True Positives (TP): 456

2) True Negatives (TN): 468

3) False Positives (FP): 32

4) False Negatives (FN): 44

The confusion matrix shows high true positive and true negative rates, with relatively low false positives and false negatives, supporting the model's robustness.

C. Comparison with Individual Models

The ensemble model's performance was compared with individual models (Gradient Boosting, Random Forest, and XGBoost). The ensemble model outperformed the individual models across all metrics:

The ensemble approach leveraged the strengths of multiple models to achieve superior accuracy, precision, recall, and F1-score.

TABLE I

MODEL EVALUATION SCORES

Model	**Accuracy**	**Precision**	**Recall**	**F1 score**	**AUC**
Gradient Boosting	89%	0.85	0.88	0.86	0.92
Random Forest	90%	0.86	0.89	0.87	0.93
XgBoost	91%	0.87	0.90	0.88	0.94
Ensemble	92%	0.89	0.91	0.90	0.95

D. Visualization of Results

Visualization tools were employed to provide insights into the model's performance and the distribution of predictions:

1) *ROC Curve:* The ROC curve demonstrated a high true positive rate and low false positive rate, confirming the model's strong discriminatory power.

2) *Precision-Recall Curve:* The precision-recall curve indicated a great level of precision and recall, emphasizing the model's balanced performance.

3) *Feature Importance:* Analysis of feature importance highlighted the key medical factors contributing to liver cirrhosis prediction, providing value.

E. Case Studies

Several in-depth case studies were conducted to demonstrate the practical application and effectiveness of the LiverLens system. These case studies involved a diverse group of patients, each with varying degrees of liver health, ranging from early-stage liver conditions to advanced liver cirrhosis. The goal was to assess the model's ability to accurately predict the onset and progression of liver cirrhosis across different patient profiles.

In each case study, the LiverLens system successfully analyzed patient data and generated predictive insights. The system not only identified those at high risk of developing cirrhosis but also provided detailed recommendations for personalized interventions. These insights proved invaluable in enabling healthcare providers to take timely and targeted actions, potentially reversing or mitigating the progression of liver disease.

The results from these case studies highlighted the model's robustness and reliability in a clinical setting, demonstrating its potential as a powerful tool for early diagnosis and proactive management of liver health. By incorporating these findings, the LiverLens system has shown that it can significantly improve patient outcomes through accurate predictions and actionable guidance tailored to individual needs.

VI. CONCLUSION

This study presents an ensemble machine learning approach to predict liver cirrhosis and provide comprehensive patient monitoring. By combining Gradient Boosting, Random Forest, and XGBoost models, we achieve high accuracy and robustness in liver cirrhosis prediction.

The platform allows patients to upload medical reports or manually enter data, facilitating early detection and timely management. Visualizations generated with Matplotlib

offer insights into liver health, aiding healthcare professionals in decision-making. Pytesseract is used for efficient text extraction, seamlessly integrating data with the MySQL database.

Future work will focus on improving model accuracy with advanced ensemble techniques, expanding the dataset, and incorporating additional features. We also plan to extend the platform to monitor other liver-related conditions, creating a comprehensive tool for liver health management.

REFERENCES

[1] P. Kuppan, N. Manoharan, "A Tentative analysis of Liver Disorder using Data Mining Algorithms J48, Decision Table and Naive Bayes," Int. J. Comput. Algorithm, vol. 6, no. 1, pp. 2278-239, 2017.

[2] A. Gulia, R. Vohra, P. Rani, "Liver Patient Classification Using Intelligent Techniques," Int. J. Comput. Sci. Inf. Technol., vol. 5, no. 4, pp. 5110-5115, 2014.

[3] Y. Kumar, G. Sahoo, "Prediction of different types of liver diseases using rule-based classification model," Technol. Health Care, vol. 21, pp. 417-432, 2013.

[4] M. Pasha, M. Fatima, "Comparative Analysis of Meta Learning Algorithms for Liver Disease Detection," J. Softw., vol. 12, no. 12, pp. 923-933, 2017.

[5] M. Abdar, N.Y. Yen, J. CS. J. Hung, "Improving the Diagnosis of Liver Disease Using Multilayer Perceptron Neural Network and Boosted Decision Trees," J. Med. Biol. Eng., vol. 4, no. 22, pp. 1-13, 2017.

[6] A.El-Shafeiy, L. Ali, E. El-Desouky, S. M. Elghamrawy, "Prediction of Liver Diseases Based on Machine Learning Technique for Big Data," Int. Conf. Adv. Mach. Learn. Technol. Appl., pp. 362-374, Springer, 2018.

[7] S. Vijayarani, S. Dhayanand, "Liver disease prediction using SVM and Naïve Bayes algorithms," Int. J. Sci. Eng. Technol. Res., vol. 4, no. 4, pp. 816-820, 2015.

[8] M. Minnoor, V. Baths, "Liver Disease Diagnosis Using Machine Learning," IEEE World Conf. Appl. Intell. Comput. (AIC), IEEE, 2022.

[9] S. Kumar, P. Rani, "A Comparative Study on Machine Learning Techniques for Prediction of Liver Disease," 6th Int. Conf. Contemp. Comput. Informatics (IC3I), IEEE, 2023.

[10] Aviral Srivastava, et al. - Automated Prediction of Liver Disease using ML Algorithms. 2022 Second ICAECT. IEEE, 2022. Cited by: Papers (6).

[11] Taher M. Ghazal, et al. - Intelligent Model to Predict Early Liver Disease using ML Technique. 2022 ICBATS. IEEE, 2022. Cited by: Papers (121).

[12] Sura Salah Rasheed, Ismaael Hadi Glob - Classifying and Prediction for Patient Disease Using ML Algorithms. 2022 3rd IT-ELA. IEEE, 2022. Cited by: Papers (1).

[13] Chappidi Aswartha Reddy, et al. - Comparative Analysis of Liver Disease Detection using ML Techniques. 2022 6th ICICCS. IEEE, 2022. Cited by: Papers (2).

[14] Lalithesh D Sawant, et al. - Analysis and Prediction of Liver Cirrhosis Using ML Algorithms. 2023 3rd CONIT. IEEE, 2023. Cited by: Papers (1).

[15] Priyadharshini K V, et al. - Leveraging Segmentation and Classification for Liver Cancer Prediction in Deep Learning. 2024 2nd AIMLA. IEEE, 2024.

[16] Tamilarasi A, et al. - Predictive Analysis for Hepatitis and Cirrhosis Liver Disease using ML Algorithms. 2022 3rd ICESC. IEEE, 2022. Cited by: Papers (4).

[17] Sonwane Suchitra Shivaji Rao, K Gangadhara Rao - Diagnosis of Liver Disease Using ANN and ML with Hyperparameter Tuning. 2024 2nd IDCIoT. IEEE, 2024.

[18] Dhriti Gada - Disease Prediction System using ML. 2022 6th ICCUBEA. IEEE, 2022. Cited by: Papers (1).

[19] Muhamamd Haseeb Aslam, et al. - Predictive Analysis on Severity of NAFLD using ML Algorithms. 2022 17th ICET. IEEE, 2022. Cited by: Papers (27).

[20] V. Saraswathi, et al. - Implementation of Hyperparameter Optimization in Liver Disease Prediction. 2022 ICPECTS. IEEE, 2022.