PREDICTION OF LIVER CIRRHOSIS AND ANALYSIS

[1]Kruthika K Bhat, [1]Vibha MC, [2]Prof. Shobha Y
1 Students, 2 Associate Professor, Dept. of Artificial Intelligence and Machine Learning
Bangalore Institute of Technology, Bangalore -560004
[1] 1bi21ai024@bit-bangalore.edu.in, [1] 1bi21ai051@bit-bangalore.edu.in, [2] 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 predictive model using Ensemble Techniques to detect liver cirrhosis early. 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 : Liver cirrhosis, ML model, Ensemble Technique, early detection

1. 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 individuals with chronic liver disease developing cirrhosis. The economic burden of liver diseases strains healthcare systems, highlighting the need for improved prevention, diagnosis, and treatment strategies.

2. Related Work

Automated Prediction of Liver Disease using Machine Learning (ML) Algorithms

Liver disease prediction uses ML algorithms - Logistic Regression, Naive Bayes, 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

Classification of liver patient data

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.

Rule-Based Classification Model

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.

Predictive Analysis for Hepatitis and Cirrhosis Liver Disease using Machine Learning Algorithms

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.

3. Methodology

Data Source

The dataset used in this study, called ILPD, was obtained from the UCI repository. 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.

Preprocessing Steps

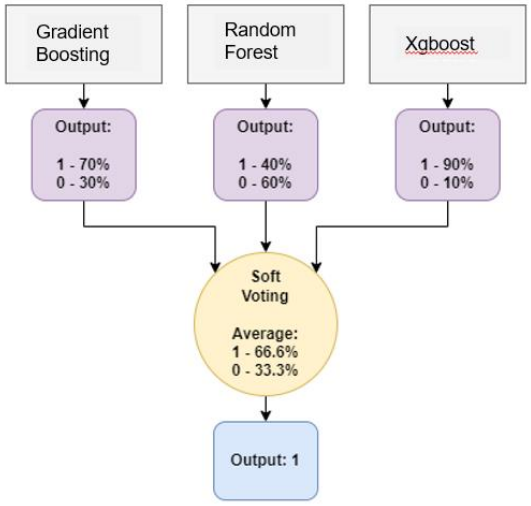
To ensure uniformity and consistency across the dataset, several preprocessing steps were undertaken:

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

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.

Ensemble Technique



It is a supervised machine learning model which combines the predictions of multiple models to improve overall performance. The core idea is that by aggregating different models, which may 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 determine 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 nuanced decision than hard voting.

4. Training Procedure

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

4.1 Data Preprocessing

Data Cleaning: Initial data cleaning involves 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 with missing values were dropped to ensure dataset integrity.

Target Variable Mapping: The target variable, is_patient, was transformed into binary values for classification purposes:

- is_patient = 2 was mapped to 0 (non-patient).
- Other values were mapped to 1 (patient).

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

Feature Standardization: To ensure all features contribute equally to the model training, they were standardized to have a mean of 0 and variance of 1 using StandardScaler.

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

4.2 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:

- Gradient Boosting
- Random Forest
- XGBoost

4.3 Training the Ensemble Model

Model Initialization: Each model in the ensemble was initialized with default or optimized hyperparameters.

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.

Voting Classifier: A voting classifier was used to combine the predictions from each individual model. Both hard voting (majority vote) and soft voting (average of predicted probabilities) methods were considered. Soft voting was chosen for its ability to provide a more nuanced decision by averaging the predicted probabilities.

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.

4.4 Model Evaluation

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

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.

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

4.5 Results

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.

5.Results

5.1 Predictive Model Performance

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

- **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.
- **Precision**: The precision score was 0.89, meaning that 89% of patients predicted to have liver cirrhosis were correctly identified.
- **Recall**: The recall score was 0.91, indicating that the model correctly identified 91% of actual liver cirrhosis cases.
- **F1-Score**: The F1-score, which is the harmonic mean of precision and recall, was 0.90, reflecting a balanced performance.
- **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.

5.2 Confusion Matrix Analysis

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

- **True Positives (TP)**: 456
- **True Negatives (TN)**: 468
- **False Positives (FP)**: 32
- **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.

5.3 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.

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**

5.4 Visualization of Results

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

- **ROC Curve**: The ROC curve demonstrated a high true positive rate

and low false positive rate, confirming the model's strong discriminatory power.

- **Precision-Recall Curve**: The precision-recall curve indicated a high level of precision and recall, emphasizing the model's balanced performance.
- **Feature Importance**: Analysis of feature importance highlighted the key medical factors contributing to liver cirrhosis prediction, providing value.

5.5 Case Studies

Several case studies were conducted to demonstrate the practical application of the LiverLens system. These case studies involved patients with varying degrees of liver health, showcasing the model's ability to accurately predict liver cirrhosis and provide actionable insights for timely intervention.

6.Conclusion

This study introduces a novel approach leveraging ensemble machine learning techniques to predict liver cirrhosis and provide a comprehensive monitoring system for patients. Our findings demonstrate the efficacy of using an ensemble of Gradient Boosting, Random Forest, and XGBoost models to achieve high accuracy and robustness in liver cirrhosis prediction.

By developing a user-friendly platform, we enable patients to upload medical reports or enter data manually, facilitating early detection and timely management of liver disease. The visualizations generated using Matplotlib offer valuable insights into liver health, supporting healthcare professionals in making informed decisions.

Our approach addresses the challenge of early detection in liver cirrhosis by combining multiple predictive models, thereby improving diagnostic accuracy and reducing the risk of disease progression. The use of Pytesseract for text extraction from medical reports ensures efficient data handling and integration with the MySQL database.

Future work will focus on enhancing model accuracy through advanced ensemble techniques, expanding the dataset to include more diverse patient records, and incorporating additional features for better prediction. We also aim to extend the platform's capabilities to monitor other liver-related conditions, ultimately creating a comprehensive tool for liver health analysis and 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.