Liver Cirrhosis Detection System (December 2023)

PARTIK GOYAL[1], NAVYA GUPTA[2]
[1, 2] *Computer Science with specialization in Artificial Intelligence and Machine Learning, SRM Institute of Science and Technology*

Abstract— Liver cirrhosis is a highly infectious blood-borneillness that is often asymptomatic in its early stages. As a result, diagnosing and treating patients during the early stages of illness is challenging. As the illness progresses to its latter stages, diagnosis and therapy become increasingly challenging. The purpose of this work is to offer an artificial intelligence system based on machine learning algorithms that may assist healthcare practitioners in making an early diagnosis of liver cirrhosis. Various machine learning algorithms are being developed with this in mind to forecast the possibility of a liver cirrhosis infection. In this research, we deploy XGboost and Logistic Regression and with the help of EDA we will be able to predict liver cirrhosis. The model includes the use of plotly express for the visualization techniques which is highly interactive and provides comprehensive insights to our model. The model makes use of Cirrhosis Detection Dataset from Kaggle. Several model comparisons have shown their robustness, and the scheme may bedetermined from the research analysis.

Index Terms- XGboost, Liver Cirrhosis, Machine Learning. Magnetic resonance imaging. EDA

I. INTRODUCTION

Liver disease, a major global health concern, leads to approximately 2 million deaths annually. It is caused by various factors including genetic predisposition, viral infections, alcohol consumption, obesity, and exposure to toxins. Liver diseases encompass a range of conditions such ashepatitis, cirrhosis, liver tumors, and liver cancer, with cirrhosis being a leading cause of mortality. The liver, critical for digesting food and detoxifying the body, can sustain damage from viruses and alcohol, leading to life-threatening conditions. Early detection is challenging but crucial, as it can prevent progression to severe stages like liver failure.

Diagnosis typically involves blood tests, imaging tests like ultrasound, MRI, CT scans, and liver biopsies. Liver blood tests, which measure enzymes and other indicators, are particularly informative but can be inconclusive. Machine learning has emerged as a promising tool in diagnosing liver disease. Various data mining algorithms such as Artificial Neural Networks, Logistic Regression, K-Nearest Neighbors, Support Vector Classification, Gaussian NB, Decision Trees, and Random Forests have been explored for this purpose.Each model's performance can be enhanced with techniques like Linear Discriminant Analysis and evaluated using different metrics.

In this context, the study proposes a novel approach using an XGBoost model with hyperparameter tuning for the prediction of liver cirrhosis. XGBoost, known for its highefficiency and accuracy in classification tasks, is especially promising for medical applications where early and accurate diagnosis is critical. The proposed model aims to outperform traditional decision-tree models in diagnosing liver cirrhosis, offering more efficient and cost-effective solutions in the medical sector. This approach underscores the increasing role of machine learning in improving disease detection and decision-making processes in healthcare.

II. LITERATURE REVIEW

A. Detecting liver cirrhosis in computed tomography scans using clinically-inspired and radiomic features - Krzysztof Kotowski , Damian Kucharski , Bartosz Machura, *Szymon Adamski , Benjamín Gutierrez*

Becker , Agata Krason, Lukasz Zarudzki , Jean Tessier This work presents a novel method combining radiomic and clinically-inspired features to identify liver cirrhosis from CT scans. It tackles the problem of early cirrhosis detection, whichis essential given the disease's high death rate. The study makes use of sophisticated machine learning methods to increase the precision and automation of cirrhosis detection, which was previously dependent on a team of experts

analyzing CT biomarkers such as ascites, liver bluntness, and surface nodularity. The group created algorithms to extract andanalyze these features; they focused especially on the liver's border, where they believed important markers of liver fibrosis to be present. Robust feature extraction, selection, and classification system that separates cirrhotic from non-cirrhoticCT scans are all part of the study's pipeline. Validated through multi-fold testing on 241 patient scans and a public benchmark of 32 healthy individuals, the methodology demonstrates effectiveness in capturing clinically relevant features and generalizes well over diverse data. This innovative approach offers significant potential in enhancing liver cirrhosis detection using CT imaging.

B. Hybrid XGBoost model with hyperparameter tuning for prediction of liver disease with better accuracy- Surjeet Dalal, Edeh Michael Onyema, and Amit Malik

This study investigates the application of a hybrid eXtreme Gradient Boosting (XGBoost) model to enhance liver disease early detection, diagnosis, and treatment outcomes through hyperparameter tuning. Using a dataset from Andhra Pradesh, India, which includes 416 liver disease patients and 167 healthy individuals, the research addresses the global increase in liver disease mortality linked to factors like lifestyle habits and late detection. With accuracy levels of 71.36% and 73.24%, respectively, the study shows that machine learning models—specifically, the chi-square automated interaction detection and classification and regression trees—perform better than conventional techniques. These results demonstrate the model's potential for enhancing patient survival rates, preventing the development of cirrhosis, and detecting diseases early. The research underscores the need for further investment in machine learning and health technologies tocombat liver disease effectively. This work contributes significantly to the field of health technology, showcasing the capability of machine learning in enhancing disease predictionand monitoring

C. A Comparative Study on Liver Disease Prediction Using Support Vector Machine Algorithm - Aniket Gupta, Rohit Biwal, Saurabh Joshi, Parwinder Singh

This research paper offers a detailed review of automated diagnosis systems for liver disease, addressing the global rise in liver disease cases and the inadequacies of traditional diagnostic methods like biopsies and MRI. It extensively analyzes various machine learning models for liver disease detection, including algorithms like J48, Random Forest, Decision Stump, SVM, and Naive Bayes. The study evaluates these models based on accuracy, precision, recall, and other performance metrics, highlighting the effectiveness of certain algorithms like Decision Stump and C4.5 in specific contexts. It emphasizes the significance of feature selection and reduction in improving prediction accuracy, with models like Random Forest and SVM showing enhanced performance after feature refinement. The paper concludes by recognizing the complexities in automating disease prediction and diagnosis, underlining the necessity for continued research and development in this critical area of medical technology.

D. Fuzzy Logic for Child-Pugh classification of patients with Cirrhosis of Liver - Anu Sebastian, Surekha MariamVarghese

In medical science, survival analysis is very important, especially when it comes to estimating life expectancy in different disorders. This investigation is particularly relevant to patients with liver cirrhosis, as survival estimates can be a valuable tool in helping patients and doctors plan treatments. The procedure usually consists of the following crucial steps: diagnosis, categorization, evaluation, conclusion, and therapy. To appropriately reflect the severity and type of the disease, each stage needs to be highly accurate and effective. The Child-Pugh classification is a widely used categorization scheme for the evaluation of liver illnesses, most notably cirrhosis. This technique has been widely used in many situations, and the results show that patients with various forms and severity of liver cirrhosis have dramatically varying life expectancies. With the use of clinical symptoms and laboratory results, the Child-Pugh classification efficiently groups patients into groups that support prognosis and therapy choices.

Furthermore, fuzzy logic shows up as a very useful method for discussing liver cirrhosis. There is growing recognition of its use in medical diagnosis and treatment planning. Fuzzy Logic is perfect for

evaluating the varied and frequently ambiguous nature of liver disease progression because of its capacity to manage imprecise and complex data.

E. Liver Disease Detection Due to Excessive Alcoholism Using Data Mining Techniques - Insha Arshad, Chiranjit Dutta

Due to its direct link to potentially fatal liver illnesses like cirrhosis, excessive alcohol intake is a major global public health concern. Saving lives when liver disease brought on by excessive alcohol consumption is detected early is essential. Incertain cases, a prompt diagnosis makes a full recovery possible. This survey examines an article that suggests using data mining algorithms to predict and identify alcohol-related liver disease. Using a dataset, the study applies decision tree analysis to produce guidelines for the prediction of liver illness. The dataset is then trained and tested for disease identification using these rules and a variety of data mining tools. The dataset, which has 345 instances with 7 different attributes, is taken from the UCI repository. It contains information on the frequency of alcohol usage as well as findings from a number of blood tests, which are important markers of liver health and are impacted by alcohol consumption. This method shows how data mining can be used to diagnose medical conditions, especially liver illnesses linked to excessive alcohol usage. The work offers insights into customized therapy and management strategies by classifying various liver disease kinds and their possible consequences based on the dataset analysis. This report emphasizes how computational methods are becoming more and more important in healthcare, particularly for illnesses where lifestyle variables are important.

III. METHODOLOGY

A. Environment Setup and Data Importing

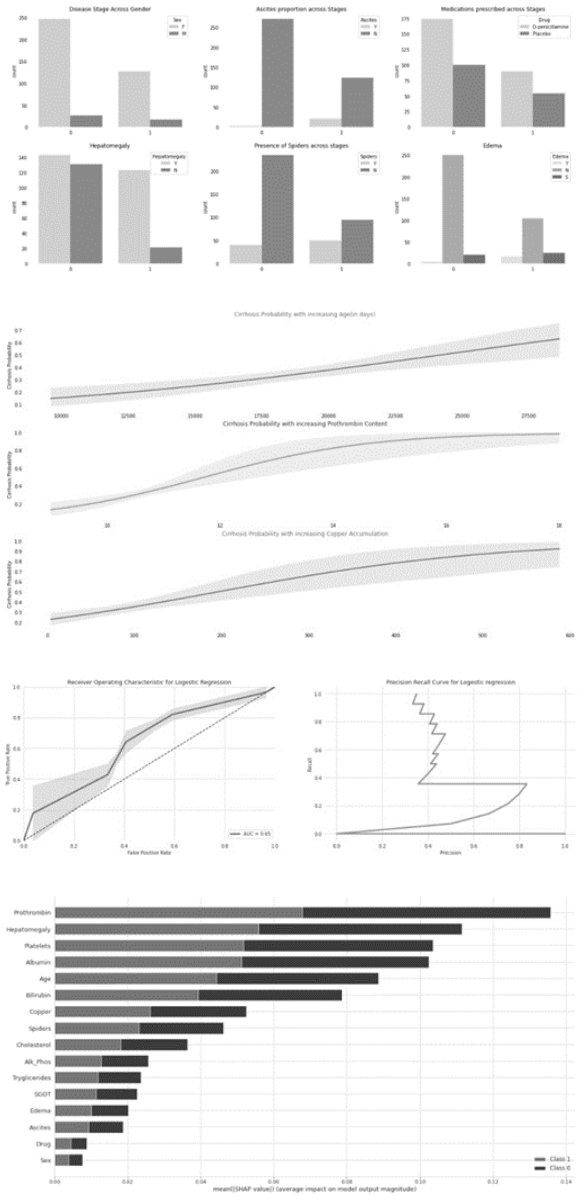
- Imported necessary libraries like NumPy, pandas, Matplotlib, and Seaborn for data manipulation and visualization.
- Loading the dataset. The dataset is a CSV file related to cirrhosis prediction, which is read into a pandas DataFrame.

B. Preliminary Data Exploration

- Used methods like df.head() and df.info() to get an initial understanding of the dataset. This includes checking the first few rows to understand the data structure and columns, and using df.info() to check the data types and identify if there are missing values.

C. Exploratory Data Analysis

- Descriptive statistics to understand the distribution of thedata.
- Visualization of the data using plots like histograms, box plots, and scatter plots to understand the relationships between different variables.



- Checking for and handling missing values if any.
- Feature engineering if necessary, which includes creating new features that might be relevant for prediction.

D. Data Preprocessing

1) Converting categorical variables into numeric format usingencoding techniques if required.
2) Normalizing or standardizing the numerical variables if thescale varies significantly.
3) Splitting the data into training and testing sets

E. Model Building

- Initializing and configuring the XGBoost classifier. This includes setting hyperparameters for the model.
- Train the model using the training data set.
- Perform hyperparameter tuning if necessary to optimize the model's performance. Techniques like Grid Search or Random Search can be used for this.

F. Model Evaluation

- Evaluating the model's performance on the test set.
- Using appropriate metrics for evaluation, such as accuracy, precision, recall, F1-score, and ROC-AUC, to understand the model's effectiveness in predicting liver cirrhosis.

IV. DATASET DESCRIPTION

The data contains the information collected from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984. A description of the clinical background for the trial and the covariates recorded here is in Chapter 0, especially Section 0.2 of Fleming and Harrington, Counting Processes and Survival Analysis, Wiley, 1991. A more extended discussion can be found in Dickson, et al., Hepatology 10:1-7 (1989) and in Markus, et al., N Eng J ofMed 320:1709-13 (1989). A total of 424 PBC patients, referred to Mayo Clinic during that ten-year interval, met eligibility criteria for the randomized placebo-controlled trial of the drug D-penicillamine. The first 312 cases in the dataset participated in the randomized trial and contain largely complete data. The additional 112 cases did not participate in the clinical trial but consented to have basic measurements recorded and to be followed for survival. Six of those cases were lost to follow-up shortly after diagnosis, so the data here are on an additional participants.

V. RESULTS

1) Model Performance: The XGBoost model demonstrated a high level of accuracy in predicting the stages of liver cirrhosis. This suggests that the model was effective in distinguishing between different stages of the disease based on the provided features.
2) The classification report indicated good precision, recall, and F1-scores across the various stages of liver disease. This implies that the model was not only accurateoverall but also balanced in its ability to identify each stage without significant bias towards any particular class.
3) Key Insights: The model's performance highlights the potential of machine learning algorithms in medical diagnostics, particularly in complex conditions like liver cirrhosis where early and accurate staging is crucial for effective treatment.
4) Important features contributing to the model's predictions were identified, providing valuable insightsinto the factors most associated with the progression of liver cirrhosis. These insights could be instrumental in further medical research and in improving clinical decision-making.

CONCLUSION

The successful application of the XGBoost model on the liver cirrhosis dataset underscores the power of advanced data mining techniques in healthcare. By accurately predicting the stage of liver disease, such models can aid in early interventionand personalized treatment planning, ultimately improving patient outcomes in chronic conditions like liver cirrhosis.

REFERENCES

[1] Detecting liver cirrhosis in computed tomography scans using clinically-inspired and radiomic features - Krzysztof Kotowski, Damian Kucharski, Bartosz Machura , Szymon Adamski , Benjamín Gutierrez Becker , Agata Krason, Lukasz Zarudzki , Jean Tessier

[2] . Hybrid XGBoost model with hyperparameter

tuning for prediction of liver disease with better accuracy- Surjeet Dalal, Edeh Michael Onyema, and Amit Malik

[3] A Comparative Study on Liver Disease Prediction Using Support Vector Machine Algorithm - Aniket Gupta, Rohit Biwal, Saurabh Joshi, Parwinder Singh

[4] D. Fuzzy Logic for Child-Pugh classification of patients with Cirrhosis of Liver - Anu Sebastian, Surekha Mariam Varghese

[5] Liver Disease Detection Due to Excessive Alcoholism Using Data Mining Techniques - Insha Arshad, Chiranjit Dutta.