RESEARCHARTICLE OPENACCESS

Statistical Machine Learning Approaches to Liver Disease Prediction

Robin Biju
Department of Computer Application,
Musaliar College of Engineering & Technology, Pathanamthitta, Kerala
The APJ Abdul kalam Technological University

Abstract:

The improvement of patient care, research, and policy is significantly impacted by medical diagnoses. Medical practitioners employ a variety of pathological techniques to make diagnoses based on medical records and the conditions of the patients. Disease identification has been significantly enhanced by the application of artificial intelligence and machine learning in conjunction with clinical data. Data-driven, machine learning (ML) techniques can be used to test current approaches and support researchers in potentially innovative judgments. The goal of this work was to use ML algorithms to derive meaningful predictors of liver disease from the medical data of 615 persons.

Keywords: - **Machine Learning, Random Forest.**

I. INTRODUCTION

The number of patients with liver disease has been steadily rising as a result of excessive alcohol use, exposure to hazardous gases, ingestion of tainted foods such pickles and cucumbers, and drug usage. In an effort to lighten the load on doctors, this dataset was used to assess prediction systems. The data set consists of the patient's age, gender, and total bilirubin. Direct bilirubin, alkaline phosphatase, alamine aminotransferase, aspartate aminotransferase, total proteins, albumin, and the ratio of albumin to globulin are other examples. Set: the field that was utilised to divide the data into two sets (patient with liver disease, or no disease).

This study attempts to find an appropriate machine learning algorithm that can determine whether a person has liver disease or not given a dataset containing biological and diagnostic data of 583 Indian patients.

A. Relevance of theproject

Using certain characteristics such as total bilirubin, direct bilirubin, alkaline phosphatase, total protein, albumin, and globulin, this software can determine whether a patient has liver disease or not.

B. Scope of theProject

It is necessary to use supervised learning to resolve this binary classification issue. Each data point has ten attributes, and there is a label that indicates if the patient has liver disease or not.In order to find the answer, our goal should be to train a variety of supervised learning models on this dataset in order to create a high-performing model that can accurately identify any new data point as positive or negative and outperform the benchmarks.

II. EXISTINGSYSTEM

Only two systems exist in the same domain, according to a thorough investigation into the subject. First, the system is entirely manual. It has the capacity to store patient information and medical records. The initial system's key characteristics are as follows. The second system is more effective than the first. It was discovered from a related research study that the system is constructed utilizing the KNN method.

.

LIMITATIONS:

□ The entire system was manual.

□ It fails to accurately predict a value using the KNN algorithm.

□ This system takes a long time to provide the user with an output.

III. PROPOSEDSYSTEM

Using the Random Forest algorithm, this system forecasts liver illness. Compare the capacity to forecast binary classifications of liver disease among several statistical learning techniques. Obtain confusion matrices for contrasting predictive classes with actual classes, then evaluate several ML techniques to gauge how well they work at diagnosing liver illness. Analyze receiver operating characteristic (ROC) curves to assess the diagnostic value of the binary liver disease classification.

The following modules make up the bulk of the proposed system:

□ RegistrationModule:

By providing the required information, the patient and hospital can register on the website using this module.

□ Login Module:

Using their username and password, registered patients, hospitals, and physicians can access the websites and their contents.

□ Liver Disease Prediction:

Use the patient's age, total bilirubin, direct bilirubin, alkaline phosphatase, total proteins, albumin, albumin, and globulin ratio to determine whether the patient has liver disease or not.

□ Donor Details:

View information about the blood types and organs used during transplantation.

IV. METHODOLOGY

In this project, we gather data from a data set, and the health specialist can enter the data for testing using our web application. In this application, we perform data cleaning and pre-processing, extensive data analysis, data visualization, and machine learning using supervised learning algorithms, decision trees, K nearest neighbor's, logistic regression, and support vector machines.This approach makes predictions about a person's liver condition based on variables including total bilirubin, direct bilirubin, albumin, total protein, etc. Additionally, the donor and recipient information that is needed for the transplantation of the liver and blood can be added to the system.

C. Random Forest Algorithm

Supervised machine learning algorithms like random forest are frequently employed in classification and regression issues. On various samples, it constructs decision trees and uses their average for classification and majority vote for regression. Bagging has been significantly modified by random forest, which creates a sizable group of de-correlated trees that may subsequently be averaged. Similar to boosting in many ways, RF is also simple to train and tune. p identical independent random variables with a variance of two are averaged. By lowering the correlation between trees without significantly raising the variance, random forest enhances the variance reduction of bagging. Take into account that for each p between 1 and P, a bootstrap sample W of size P from the training data is possible.The bootstrapped data can then be used to grow the random forest tree Tp. The technique is then repeated for each terminal node of the tree until the minimal node size nmin is reached. This splits the node into two daughter nodes and distributes m variables at random from the p variables. Finally, the ensemble of trees can be discovered by providing the sequence "Tp 1 P." The forecast for a fresh point x is given. Consequently, Cp (x) is the pth random forest tree's class forecast for classification.

V. LITERATUREREVIEW

D. Software-based Prediction of Liver Disease with Feature Selection and Classification Techniques. [JagdeepSingh 1970–1980]

Today, everyone's health is a very essential concern, so it is necessary to offer medical services that are freely accessible to everyone. The primary goal of this study is to forecast liver illness using a software engineering methodology that makes use of feature selection and classification techniques. The Indian Liver Patient Dataset (ILPD) from the University of California, Irvine database is used to carry out the proposed research. The many variables of the liver patient dataset, including age, direct bilirubin, gender, total bilirubin, Alkphos, sgpt, albumin, globulin ratio, and sgot, among others, are used to forecast the risk level of liver illnesses.

On the Liver Patient dataset, several classification techniques are applied to determine accuracy, including Logistic Regression, Sequential Minimal Optimization, and K-Nearest Neighbor.

E. Biochemical Evaluation of Patients of Alcoholic Liver Disease and Non-alcoholic Liver Disease.[PRASAD.P.TORKADI 1979–1983]

The fundamental drawback of this approach is that, while the KNN algorithm predicts the outcome with a moderate degree of accuracy, it classifies the data according to the dataset's majority. Alcohol abuse over an extended period of time causes alcoholic liver disease (ALD). It might be challenging to distinguish ALD from non-ALD (non-alcoholic steatohepatitis, viral hepatitis), as the patient may deny drinking. Because ALD patients are managed differently than individuals without ALD, accurate diagnosis is crucial. This system's objectives were to (1) compare the biochemical parameters of ALD and non-ALD patients to controls, and (2) determine whether these parameters can distinguish between ALD and non-ALD.The study involved 35 patients with acute viral hepatitis and 50 patients with alcoholic liver disease (ALD) in groups I and II, respectively. Our research shows that serum AST/ALT ratio, GGT, and ALP measurements may reliably distinguish ALD patients from NASH and acute viral hepatitis.

F. Liver Disease Prediction using Naïve Bayes Algorithms. [Dr. S. Vijayarani 1816–1820]

Data mining has recently improved the simplicity of use for disease prediction in the healthcare sectors. The process of extracting information from huge datasets, warehouses, or other repositories is known as data mining. Predicting diseases using the vast medical datasets is an extremely difficult task for academics. The researchers employ data mining techniques including classification, clustering, association rules, and others to address this problem. This study's primary goal is to use classification algorithms to predict liver disorders. Naive Bayes algorithms were employed in this study. Based on their performance characteristics, such as classification accuracy and execution time, these classifier algorithms are contrasted.

G. Evaluation of Abnormal Liver Tests [Tinsay A. Woreta 2014]

The diagnosis and treatment of liver illnesses both heavily rely on the use of serum biochemical testing. The routine use of such tests has boosted the diagnosis of liver illnesses in patients who would not otherwise exhibit any symptoms, frequently offering the first indication of liver pathology. In most circumstances, these laboratory tests can assist clinicians in identifying the cause of liver illness in addition to a thorough history, physical examination, and imaging studies. Based on the degree of aminotransferase increase relative to alkaline phosphatase, liver damage has traditionally been classified as mostly hepatocellular or cholestatic. There is frequently significant overlap in the presentation of different liver disorders, which frequently have a mixed pattern, despite the fact that such a differentiation might help orient early evaluation.

VI. CONCLUSION

Clinicians who are skilled at identifying noteworthy observations and categorising them as normal or abnormal using background knowledge and other context clues can detect chronic liver disease. Similar to how ML algorithms may help medical professionals, these algorithms can be trained to recognise the potential for liver illness. ML approaches were able to distinguish between blood donors with and without liver disease with high accuracy by using the correlation of each variable with the risk of liver disease to train the model. By increasing awareness of risk factors and diagnostic variables, the application of ML approaches can aid in lowering the overall burden of liver disease on public health globally.More importantly, for chronic liver illness, ML could reduce liver-related mortality, transplants, and/or hospitalizations by identifying liver disease in its early stages or in concealed cases.

REFERENCE

[1] Asrani, S.K.; Devarbhavi, H.; Eaton, J.; Kamath, P.S. "Burden of liver diseases in the world". J. Hepatol. 2019.

[2] Chalasani, N.; Younossi, Z.; Lavine, J.E.; Charlton, M.; Cusi, K.; Rinella, M.; Harrison, S.A.; Brunt, E.M.; Sanyal, A.J. "The diagnosis and management of nonalcoholic fatty liver disease: Practice guidance from the American Association for the Study of Liver Diseases". Hepatology 2018.

[3] Wang, Y.; Li, Y.; Wang, X.; Gacesa, R.; Zhang, J.; Zhou, L.; Wang, B. "Predicting Liver Disease Risk Using a Combination of Common Clinical Markers: A Screening Model from Routine Health Check-Up". Dis. Markers 2020.