# STATISTICAL MACHINE LEARNING APPROACHES TO LIVER DISEASE PREDICTION

## TEAM ID: PNT2022TMID50136

## TEAM MEMBERS: S.Gayathri,V.Jemila Devi,N.Karthika,P.Sneka.

**Paper 1:** Fahad Mostafa , Easin Hasan , Morgan Williamson and Hafiz Khan, Statistical Machine Learning Approaches to Liver Disease Prediction, 1 December 2021 Published.

Using machine learning algorithms to predict disease is made possible by increasing access to hidden attributes in medical data sets. Various kinds of data sets, such as blood panels with liver function tests, histologically stained slide images, and the presence of specific molecular markers in blood or tissue samples, have been used to train classifier algorithms to predict liver disease with good accuracy. The ML methods described in previous studies have been evaluated for accuracy by a combination of confusion matrix, receiver operating characteristic under area under curve, and k-fold cross-validation. Singh et al. designed software based on classification algorithms (including logistic regression, random forest, and naive Bayes) to predict the risk of liver disease from a data set with liver function test results . Vijayarani and Dhavanand found that SVM performed better over naive Bayes to predict cirrhosis, acute hepatitis, chronic hepatitis, and liver cancers from patient liver function test results . SVM with particle swarm optimization (PSO) predicted the most important features for liver disease detection with the highest accuracy over SVM, random forest, Bayesian network, and an MLP-neural network. SVM more accurately predicted drug-induced hepatotoxicity with reduced molecular descriptors than Bayesian and other previously used models . Phan and Chan et al. demonstrated that a convolutional neural network (CNN) model predicted liver cancer in subjects with hepatitis with an accuracy of 0.980 . The ANN model has been used to predict liver cancer in patients with type 2 diabetes. Neural network ML methods can help differentiate between types of liver cancers when applied toimaging data sets . Neural network algorithms have even been trained to predict a patient's survival after liver tumor removal using a data set containing images of processed and stained tissue from biopsies . ML methods can facilitate the diagnosis of many diseases in clinical settings if trained and tested thoroughly. More widespread application of these methods to varying data sets can further improve accuracy in current deep learning methods. This study aimed to (i) impute missing data using the MICE algorithm; (ii) determine variable selection using eigen decomposition of a data matrix by PCA and to rank the important variables using the Gini index; (iii) compare among several statistical learning methods the ability to predict binary classifications of liver disease; (iv) use the synthetic minority oversampling technique (SMOTE) to oversample minority class to regulate overfitting; (v) obtain confusion matrices for comparing actual classes with predictive classes; (vi) compare several ML approaches to assess a better performance of liver disease diagnosis; (viii) evaluate receiver operating characteristic (ROC) curves for determining the diagnostic ability of binary classification of liver disease.

**Paper 2:** Rakshith D B ,Mrigank Srivastava ,Ashwani Kumar ,Gururaj S P, #Department of Computer Science and Engineering, Siddaganga Institute of Technology, Tumkur,India, Liver Disease Prediction System using Machine Learning Techniques. Vol. 10 Issue 06, June-2021.

Naive Bayes

Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. They are among the simplest Bayesian network models, but coupled with kernel density estimation, they can achieve higher accuracy levels. Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximumlikelihood training can be done by evaluating a closedform expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

SVM

Support Vector Machine or SVM algorithm is a simple yet powerful Supervised Machine Learning algorithm that can be used for building both regression and classification models. SVM algorithm can perform really well with both linearly separable and non-linearly separable datasets. Even with a limited amount of data, the support vector machine algorithm does not fail to show its magic. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. In short, the hyperplane is (n-1)- D plane for n features.

PyQt Library

PyQt is a GUI widgets toolkit. It is a Python interface for Qt, one of the most powerful, and popular crossplatform GUI library. PyQt is a blend of Python programming language and the Qt library. PyQt API is a set of modules containing a large number of classes and functions. While QtCore module contains non-GUI functionality for working with file and directory etc., QtGui module contains all the graphical controls. In addition, there are modules for working with XML (QtXml), SVG (QtSvg), and SQL (QtSql), etc. For this paper, we have used the PyQt version 5, which is implemented as more than 35 extension modules and enables Python to be used as an alternative application development language to C++ on all supported platforms including iOS and Android.

Spyder Notebook

Spyder is an open-source cross-platform integrated development environment (IDE) for scientific programming in the Python language. Spyder is extensible with first-party and third-party plugins, includes support for interactive tools for data inspection and embeds Python-specific code quality assurance and introspection instruments, such as Pyflakes, Pylint and Rope. It is available cross-platform through Anaconda, on Windows, on macOS through MacPorts, and on major Linux distributions. Spyder uses Qt for its GUI and is designed to use either of the

PyQt or PySide Python bindings. QtPy, a thin abstraction layer developed by the Spyder project and later adopted by multiple other packages, provides the flexibility to use either backend.

**Paper 3:** Robin Biju Department of Computer Application, Musaliar College of Engineering & Technology, Pathanamthitta, Kerala The APJ Abdul kalam Technological University, Statistical Machine Learning Approaches to Liver Disease Prediction, Volume 5 Issue 4, July-August 2022, Available at www.ijsred.com

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.

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 (nonalcoholic 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.

**Paper 4:** Vijay Panwar, Naved Choudhary, Sonam Mittal, Gaurav Sahu, REVIEW OF LIVER DISEASE PREDICTION USING MACHINE LEARNING ALGORITHM, © 2021 JETIR February 2021, Volume 8, Issue 2.

Health care and medicine handles huge data on daily basis. This data comprises of information about the patients, diagnosis reports and medical images. It is important to utilize this information to decipher a decision support system. To achieve this it is important to discover and extract the knowledge domain from the raw data. It is accomplished by knowledge discovery and data mining (KDD) [3]. The implementation of data mining techniques is widespread in biological domain. In recent years, liver disorders have excessively increased and liver diseases are becoming one of the most fatal diseases in several countries. In this study, liver patient datasets are investigate for building classification models in order to predict liver disease. Several feature model construction and comparative analysis are implemented for improving prediction accuracy of Indian liver patients. Different studies have been conducted for classification of liver disorders, they are discussed briefly. Classification algorithm is one of the greatest significant and applicable data mining techniques used to apply in disease prediction. Classification algorithm is the most common in several automatic medical health diagnoses. Many of them show good classification accuracy. In another study the UCI liver dataset was used for selection of sub features based on random forest classifier with multi-layer perceptron induced [4].Different approaches for artificial intelligence for the liver patient dataset, precise predictions of liver failure were applied [5, 6, 7 and 8]. Identification of liver infection at preliminary stage is important to combat the frequency and severity deaths of patients in India. The patients must be screened based on initial symptoms for development of personalized therapy. In this study, an attempt is made for prediction of liver disease in patients using data mining techniques. Based on the review of literature, it was depicted that the past research studies have implemented different data mining techniques for classification of liver dataset. A hybrid model can be adapted to further increase the prediction accuracy of liver disease. It is followed by development of a graphical user interface would further aid the scientific community in early diagnosis of liver infection. It will provide a framework for end user application for generating promising treatment protocols.