Prediction and Diagnosis of Liver Disease Using Machine Learning Models

Mr. Narendra G [1] , Mr. Tejas SV [2] , Mr. Vishnu Teja S Hingoli [3] , Mr. Pradeep E [4] , Mr. Narayana H M [5]
[1,2,3,4] *Students, Department of Computer Science and Engineering , M.S Engineering College, Bangalore, India*
[5] *Associate Professor, Department of Computer Science and Engineering , M.S Engineering College, Bangalore, India*

***Abstract*— Liver disease is one of the key causes of high numbers of deaths in the country and is considered a life-threatening disease, not just anywhere, but worldwide. Liver disease can also impact peoples early in their life. More than 2.4 percent of annual Indian deaths are due to liver disorders. It is also difficult to detect liver disease due to mild symptoms in the early stages. If it is too late the signs always come to light. Thus liver-related disease poses more problems for people living and is more important nowadays to recognize the causes, and identification phase. So, for early detection of liver disease, an automated program is needed to build with more accuracy and reliability. Specific machine learning models are developed for this purpose to predict the disease. In this paper, the methods of Support Vector Machines (SVM), Decision Tree (DT) Neural Network and Random Forest (RF) is proposed to predict liver disease with better precision, accuracy and reliability.**

Keywords: Mean Square Error (MSE), Mean Absolute Error (MAE), R-Squared Error, Root Mean Square Error (RMSE)

INTRODUCTION

Liver infection is a precarious malady to analyze given the nuance of the indications while in the beginning periods. Issues with liver illnesses are not found until it is regularly past the point of no return as the liver keeps on working in any event, when incompletely harmed. Early determination can conceivably be life sparing. While it is difficult to diagnose even the experienced herbal practitioner, the early signs of these diseases that be identified. Late patient experiences will graciously build up his / her standards of life. Therefore, the findings of this analysis are important from the point of view of both the computer scientist and the medical professional. This project proposes the three machine learning models, i.e., help for Vector Machines, Decision Tree and Random Forest and Neural Network for the prediction of liver disease.

Liver disease is one of the key causes of high numbers of deaths in the country and is considered a life-threatening disease, not just anywhere, but worldwide.

Liver disease can also impact peoples early in their life. More than 2.4 per cent of annual Indian deaths are due to liver disorders. It is also difficult to detect liver disease due to mild symptoms in the early stages. If it is too late the signs always come to light.

Thus liver-related disease poses more problems for people living and is more important nowadays to recognize the causes, and identification phase.

So, for early detection of liver disease, an automated program is needed to build with more accuracy and reliability.

Specific machine learning models are developed for this purpose to predict the disease.

EXISTING SYSTEM

Vasan Durai [1] et al presented a liver disease prediction based on the machine learning model. The proposed supervised classification algorithm J48 algorithm for prediction. The dataset is collected from the UCI repository dataset.

Nazmun [2] et al proposed liver disease classification by using decision tree classifier. It achieved higher accuracy than other algorithms.

Ramalingam [3] et al proposed the machine learning model for prediction of liver-related diseases. For this work, liver data was collected from UCI which is related to hepatitis and hepatocellular carcinoma liver disease. They proposed the K means algorithm, SVM, LR, RF and Neural Networks for liver disease prediction.

Joel Jacob [4] et al presented a novel approach for liver disease detection. The Indian Liver Patient Dataset (ILPD), picked from the learning repository of UCI, was retrieved for this analysis. They implemented the Regression model, KNN model and ANN model liver disease detection.

Shambel [5] et al in proposed a data mining technique to predict and analyze the liver disorder. They used SVM, decision tree classifiers. They performed data partitioning based on test set to test the model. Dataset repository UCI is used for prediction. Comparison of the type of liver disease discussed.

Ignisha Rajathi [6] et al proposed the hybrid WOA-SA and ensemble classifier was proposed. They proposed the method got better accuracy, sensitivity and specificity.

A. Disadvantages of Existing System

- Accuracy is less.
- Huge data is required.
- Regression techniques are providing accuracy less than 75%.

PROPOSED SYSTEM

This project presents the three distinct machine algorithms i.e., SVM, DT, and RF to predict or classify the liver disease by using different attributes. The proposed method for liver disease prediction, as follows

1. UCI machine learning library gathers data from the liver disease dataset.
2. The dataset has some -1 values; numerical values are required to replace it. This process is carried out in the pre-processing phase.
3. Our data set is divided into data for validation for the training and testing process.
4. Finally trained data can be evaluated with various algorithms and test data are graded based on a trained model.

SYSTEM ARCHITECTURE

Fig 1: System architecture

The above figure 1 illustrates the steps in liver disease detection, the dataset is download from UCI machine learning algorithm. Then we will be preprocessing and split the data into training data and testing data then applying the machine learning algorithm to classify the liver disease and shows the performance evaluation.

METHODOLOGY

SUPPORT VECTOR MACHINE

SVM is one of the most widely used computer supervised learning model to perform prediction and classification. SVM has used the hyperplane in the function space that differentiates the labels or groups. An SVM model interprets the training data points as points within the function domain, distributed in a way that distinguishes as broadly as possible the points belonging to different classes the same area, the test data points are then drawn and graded by which side of the threshold they fall.

RANDOM FOREST

Random Forest is an excellent supervised learning algorithm that can train a model to predict which classification results in a certain sample type belong to base on a given dataset's characteristic attributes and classification results. Random Forest is based on a decision tree and adopts the Bagging (Bootstrap aggregating) method to create different training sample sets. The random subspace division strategy selects the best attribute from some randomly selected attributes to split internal nodes. The various decision trees formed are used as weak classifiers, and multiple weak classifiers form a robust classifier, and the voting mechanism is used to classify the input samples. After a random forest has established a large number of decision trees according to a certain random rule when a new set of samples is input, each decision tree in the forest makes a prediction on this set of samples separately, and integrates the prediction results of each tree, get a final result.

DECISION TREE

Decision Tree is one of the Supervised Learning algorithms. Classification issues are mostly dealt with by the use of a decision tree. It works easily with attributes which are constant and categorical. Based on important predictors, the population divides into two or more related DT sets. The DT's first step is to calculate entropy for each and every attribute. Next, the dataset is split with high information gain or less entropy based on the variables/predictors. Remaining attributes are followed by two steps above.

$$Entropy\ (E) = \sum_{k=1}^{i} -q_k \log_2 q_k \qquad \text{....Equ (1)}$$

Where *l* the variable response modules count is referred to q_k, is the ratio of the number of k^{th} class methods to the total number of models

NEURAL NETWORK

Neural Networks are computing system with interconnected nodes that work much like neurons in human brain. Using algorithms, they can recognize hidden patterns and correlation raw data, cluster and classify it.

Neural network is also ideally suited to help solve complex problems in real life situations. They can learn and model the relationships between input and outputs that are

nonlinear and complex, make generalizations and inferences, reveal hidden relationships, patterns and predictions. As a result, neural networks can improve decision process in many areas. For example

- Medical and Disease diagnosis
- Financial predictions for stock prices, currency etc.
- Credit card and medical fraud detection and many more.

IMPLEMENTATION

Dataset Collection:
UCI machine learning library includes the Patient Dataset for Indian Liver. The data set includes 416 reports of liver disease patients and 167 records of nonliver disease patients. The data collection was collected from Andhra Pradesh in north-eastern India. The data collection contains 441 records for males' patients and 142 records for female's patients. 10 attributes are being used such as age of the person, gender, total bilirubin, direct bilirubin, direct bilirubin, alkaline phosphotase, alamine, aspartate aminotransferase, total protein, albumin and albumin ratio, and the liver disease prediction globulin ratio.

Data Pre-processing:
The dataset is composed of original attributes and -1 values. The -1 values cannot be processed in programming This transforms these values into a different value, i.e. a numerical value. -1 Column mean values are replaced.
Pseudo code:
Procedure preprocess()
Input: Dataset
Output: Cleaned data
Begin
Step 1: Load the dataset
Step 2: Read dataset using dataframe
Step 3: find -1 in the dataset
Step 4: If any values ==-1 then
Calculate mean value using np.mean(column)
Step 5: Replace mean value with -1
Return cleaned data

Data Splitting:
Throughout data slicing, the entire dataset is divided into data for the training and testing of the process analysis. In that, 80 percent of data is used for demo data for training and 20 percent for testing.

Pseudo code:
Procedure Split()
Input: cleaned dataset
Output: trainset, testset
Begin
Step 1: Load the dataset
Step 2: Read dataset using dataframe
Step 3: Split the data set into train set 75% and test set 25% using sklearn
Return Trainset and testset

Classification:
In classification, accurate detection of disease by using training and testing data set. It proposes three models of machine learning to build the prediction. First training data is trained across three models of machine learning i.e. Decision Tree, Random Forest and Support Vector Machine are predicted on the basis of a trained model of learning, one by one, and then test data. Some parameters including accuracy, precision, and recall are finally compared with the above algorithms. The details given below are three different algorithms.

Pseudo code:
Procedure Train()
Input: train set, test set
Output: Trained model
Begin
Step 1: Read Train set and test set
Step 2: Build Neural network model using tensorflow
Step 3: Train the model using fit()
Step 4: Performance Graph
Return Trained Model

EXPERIMENTAL RESULTS

In our tests, we looked at the performance of training set that had different features in it.
The proposed system requires the operating system of windows 7 and above. It requires 8Gb of RAM and 120Gb of storage disk. The proposed system is coded using python programming language along with anaconda tool and some library functions.

Fig 2: Overview of the Proposed system
The above fig 2 represents the overview of the developed proposed system.

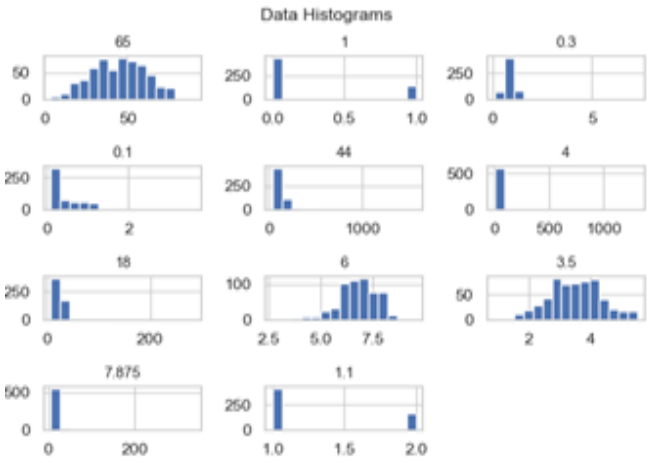

Fig 3: Data Histogram of Support Vector Machine

The above fig 3 shows the graph of data classification histogram.

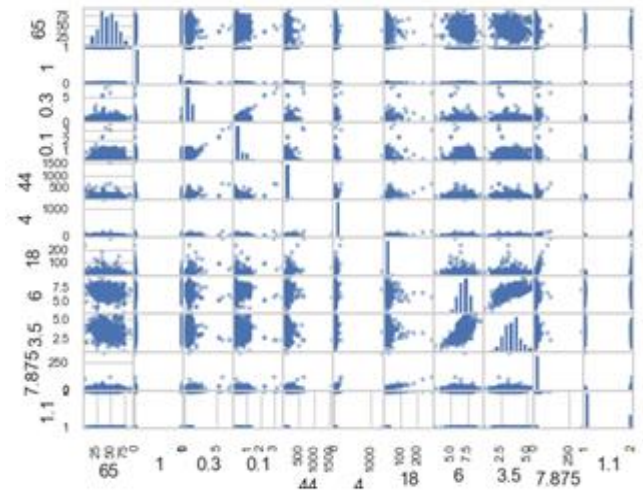

Fig 4: Resultant graph after applying SVM

The above diagram fig 4 shows the graph after plotting the graph of scattered data.

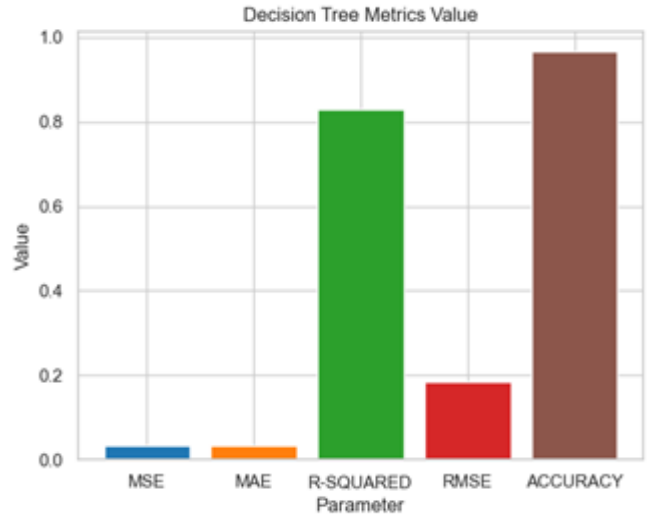

Fig 5: Resultant of Random Forest Metrix values

The above fig 5 shows the values of different attributes i.e., MSE, MAE, R-Squared parameter, RMSE used in Random Forest

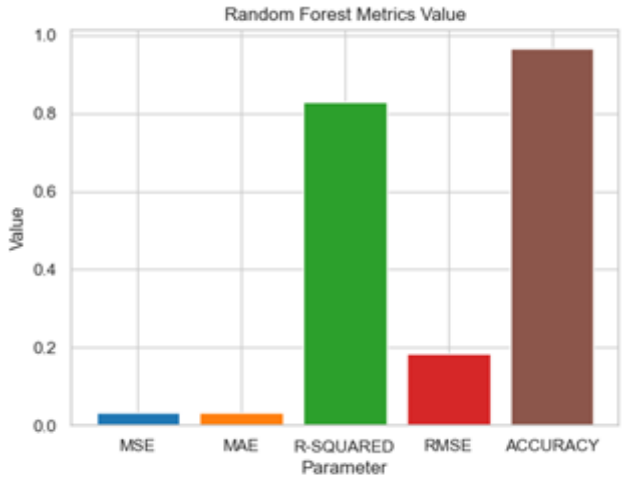

Fig 6. Resultant Graph of Decision Tree Metrix Value

The above fig 6 shows the values of different attributes i.e., MSE, MAE, R-Squared parameter, RMSE used in Decision Tree.

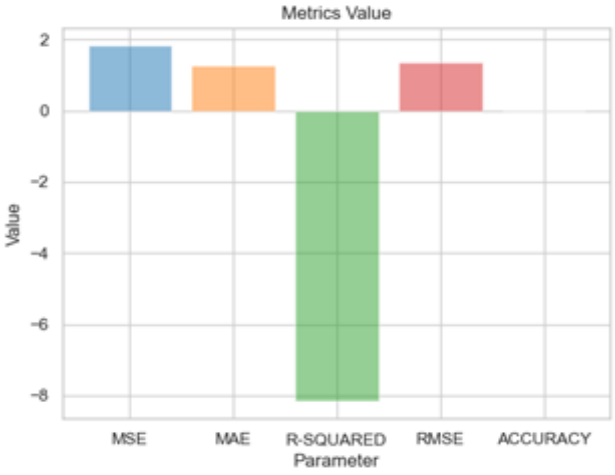

Fig 7: Resultant Graph of Neural Network Metrics Value

The above fig 7 shows the different attributes i.e., MSE, MAE, R-Squared parameter, RMSE used in Neural Network

PERFORMANCE EVALUATION

This phase involves the performance and comparative evaluation of accuracy algorithms. The performance is measured on the accuracy of the algorithm.

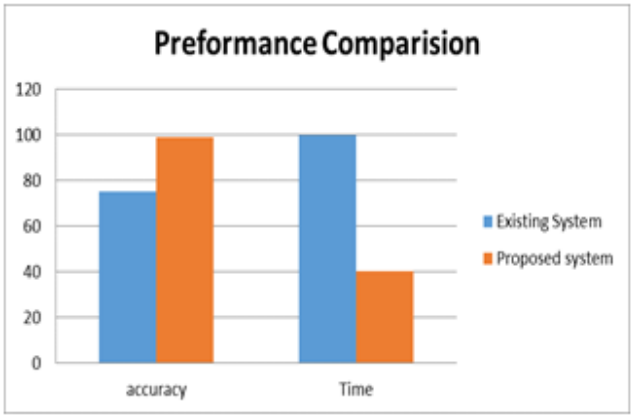

Fig 8: Performance Comparison

The above fig 8 shows the comparison graph between the Currently Existing system and Proposed System.

CONCLUSION

In this work, the different machine learning algorithms is evaluated for the prediction of liver disease. Due to the subtle nature of its symptoms, liver disease is particularly difficult to diagnose. Liver disease prediction followed the step of preparing data in that data was collected from the public database preprocessing of data for -1 value replacement. Data division into the entire array of data split into training and research. Eventually, quantitative measurement metrics such as precision, accuracy and recall are measured over various machine learning models.

FUTURE ENHANCEMENT

In future we can apply the efficient feature selection algorithms and deep learning algorithms to improve the accuracy in liver disease classifications.

REFERENCES

[1] Vasan Durai, Dinesh, Kalthireddy, "Liver disease prediction using machine learning", International Journal of Advance Research, Ideas and Innovations in Technology, 2017

[2] Nazmun Nahar and Ferdous Ara, "Liver disease prediction by using different Decision tree techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.8, No.2, March 2018

[3] V.V. Ramalingam, A.Pandian, R. Ragavendran, "Machine Learning Techniques on Liver Disease A Survey", International Journal of Engineering & Technology, 7 (4.19) (2018) 485-495

[4] Joel Jacob, Joseph Chakkalakal Mathew, Johns Mathew, Elizabeth Issac, "Diagnosis of Liver Disease Using Machine Learning Techniques" International Research Journal of Engineering and Technology (2018)

[5] Shambel Kefelegn, Pooja Kamat, "Prediction and Analysis of Liver Disorder Diseases by using Data Mining Technique: Survey", International Journal of Pure and Applied Mathematics 2018

[6] L. Alice Auxilia, "Accuracy Prediction using Machine Learning Techniques for Indian Patient Liver Disease", Proceedings of the 2nd International Conference on Trends in Electronics and Informatics (ICOEI 2018)

[7] Ain Najwa Arbain, B. Yushalinie Pillay Balakrishnan, "A Comparison of Data Mining Algorithms for Liver Disease Prediction on Imbalanced Data", International Journal of Data Science and Advanced Analytics, 2018

[8] G. Ignisha Rajathi 1,* and G. Wiselin Jiji 2, "Chronic Liver Disease Classification Using Hybrid Whale Optimization with Simulated Annealing and Ensemble Classifier", MDPI Journal 2019.

[9] Hemanth Kumar and Sivasangari.A (2016), " An Efficient Distributed Data processing Method for Smart environment" Indian Journal of Science and Technology, ISSN (Print) : 0974-6846, Vol. 9, Issue 31,pp. 1-4.

[10] P.Ajitha, A. Sivasangari, K.Indira,"Predictive Inter and Intra Parking System",International Journal of Engineering and Advanced Technology (IJEAT), ISSN: 2249 – 8958, Vol.8, Issue-2S, December 2018.

[11] Indira K, Christal Joy E, "Energy Efficient IDS for ClusterBased VANETS",Asian Journal of Information Technology, vol 14(1) ,37-41,2015.

[12] E. Brumancia, S. Justin Samuel, R. M. Gomathi and Y. Mistica Dhas,"An Effective Study on Data Fusion Models in Wireless Sensor Networks ", ARPN Journal of Engineering and Applied Sciences, Vol. 13, No. 2, January 2018.