IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

System for diagnosis of Liver Disease Using Machine Learning Technique

[1]Pragati Bhagat, [2]Sonali Meshram, [3]Rohini Ganorkar,[4]Mohit Kewate,[5]Prof. Ashish Palandurkar,[6]Ashish Gour

Department of Computer Science and Engineering,

Nagpur Institute of Technology, Nagpur, Maharashtra, India

Abstract: Many people suffer from liver disease but they don't have an idea about it. It is difficult to diagnosis of liver disease at high level. Before treatment of liver disease doctors first diagnose whether patient has liver disease or not, basis on different parameter. The system for diagnosis of liver disease using machine learning algorithms is an initiative towards better diagnosis of this disease as early as possible. Various algorithms are being studied in order to select the best algorithm which can give the best accurate results. According to the four parameters- Accuracy, Precision, Sensitivity and Specificity the algorithm is being selected. After the study ANN algorithm turned out to be the best algorithm to implement and provides more accuracy than other algorithms. So, ANN is implemented using MATLAB platform and the user interface is also constructed with the help of MATLAB.

Index Terms **– Artificial Neural Network, Liver Disease, Machine Learning.**

I. INTRODUCTION

Liver may be a crucial organ that performs many tasks related to metabolism, energy storage, and detoxification of waste. It helps to digest food, convert it to energy, and store the energy until we would like it. It also helps filter harmful substances from our bloodstream. disease may be a general term that refers to any condition affecting liver. These conditions may develop for various reasons, but they're going to all damage liver and impact its function. There are different sorts of disease. Disease caused by viruses, like hepatitis A, hepatitis B, and hepatitis C. Diseases caused by drugs, poisons, or an excessive amount of alcohol. Patient can suffer from acute disease when his liver pack up within short time, days and week. liver disease is when many fat reposes on the liver. Machine learning is that the technique that provides computer the power to find out without being explicitly programmed. There are different machine learning algorithms which are beneficial for determine the liver patient. In hospital, doctor check the liver patient on basis of the different parameter which are used to find out the liver patient from healthy individual. But its long process and it take excessive amount of time.

Patients who suffer from liver disease are continuously increasing due to high consumption of alcohol, harmful gases, intake of contaminated food, pickles and medicines. There are three preliminary steps that serve as the elementary foundation of any medical treatment paradigm. The diagnosis stage, the classification stage, the assessment stage, the conclusion stage and finally the treatment stage. All these stages are expected to be accurate to the parameters and effective in their measure to distinctly reflect the quantified magnitude and the intensity of the study of the disease in the context. Some of the machine learning approaches that are being considered, do not stand viable for a large volume of data.

The due consideration is given to the process, owing to fact that the methodology suits the conditions where there are smaller volumes of data. There are certain methodologies that are incompatible and non-cohesive when it comes to the collection of real-time data and the implementation procedures of the same. Automatic classification tools may reduce burden on doctors. This system evaluates the selected classification algorithms for the classification of some liver patient datasets. The classification algorithms considered here are support vector machine (SVM)classifier, Back propagation Neural Network algorithm, and k-nearest neighbor. These algorithms are evaluated supported four criteria: Accuracy, Precision, Sensitivity and Specificity. Graphical user interface (GUI) is designing on MATLAB, where patient can insert their data. User can communicate with the system with the help of GUI.

II. DATASET USED

The dataset used is The Indian Liver Patient Dataset (ILPD) which was selected from UCI Machine learning repository for this study. It is a sample of the entire Indian population collected from Andhra Pradesh region and comprises of 585 patient data. Machine learning with its applications emerge as remarkable methods used frequently in diagnosis of various diseases with quite successful results. Traditionally, liver disease can be diagnosed clinically by analyzing the levels of enzymes in the blood. In this research work, artificial

neural network classifier algorithm is used for predict the liver disease. After studied all algorithm , we find out that artificial neural network provide high accuracy than other algorithm.

Classification algorithms are considerably suitable and utilized in different automated diagnosis tools. pressure, eating habits, habitual drinking alcohol, exposure to toxic gas, and nature of drug intake etc. are the most causes of rising disease patients per annum. Automatic classification algorithms are often applied during this case. It reduces the line of patient at the liver experts. The Support vector machine (SVM) gives 80.78% accuracy, K-nearest neighbor (KNN) gives 78.89% accuracy, linear regression provides 75.67% accuracy and logistic regression provide 82% accuracy. Artificial neural network (ANN) gives 94.89% accuracy. So, ANN provide high accuracy than other algorithms.
The main objective of this research work was to seek out the simplest classification algorithm in terms of precision, accuracy, specificity and sensitivity. Therefore, the present investigation was done to determine the relative performance of five classification algorithms namely, Support vector machine (SVM), K-nearest neighbor (KNN), Artificial neural network (ANN), logistic regression and linear regression. The dataset contains 583 records with 10 features means including Age, Gender, Total Bilirubin,Direct Bilirubin, Alkaline Phosphatase, Alamine Phosphatase, Total Proteins, Albumin, Albumin and Globulin Ratio, result. The dataset was split into two sets that included 486 records of group 1(liver patients) and 100 records for group 2(non-liver patients).

III. EXPERIMENTAL SETUP

The liver patient are classifies from non liver patient by using different parameters. There are ten parameter are use to predict the disease . the outline of parameters are as follows:

Age: Age of patient

Gender: Gender of patient.

Total bilirubin : A term that any form of a yellowish pigment made in the liver when red blood cells are broken down and normally excreted with the bile; also termed "total bilirubin." Bilirubin can be classified as indirect while it is circulating in the body.

Direct Bilirubin: Indirect *bilirubin* is formed by the breakdown of hemoglobin in the red blood cells. The liver converts this *bilirubin* into *direct bilirubin*, which can then be released into the intestine by the gallbladder for elimination.

Alkaline phosphatase: It is also called Alk Phos. It is an enzyme (EN-zeyem) found in your bones and liver. An enzyme is something that helps speed up a chemical reaction in your body. This blood test measures the amount of alkaline phosphatase in the blood.

Alamine Phosphatase: It is an odorless, white crystalline solid which is often used in liquid or gel form. It is used in ceramics, dental cements, cosmetics, paints, paper and pharmaceuticals. **Total protein :** The total protein test measures the total amount of two classes of proteins found in the fluid portion of your blood. These are albumin and globulin. Proteins are important parts of all cells and tissues.

Albumin and globulin: Normal range of albumin is 39 to 51 grams per liter (g/L) of blood. The normal range for globulins varies by specific type. A normal range for total globulins is 23 to 35 g/L.

Albumin : It is a protein made by your Liver. It helps keep fluid in our bloodstream so it doesn't leak into other tissues. It is also carries various substances throughout our body, including hormones, vitamins, and enzymes. Low albumin levels can indicate a problem with our liver or kidneys.

IV.WORKING OF MACHINE LEARNING ALGORITHM:

The machine learning component is considered to be the brain of the system where all the learning aspects take place and are controlled centrally. The machine learning algorithms enable the system to learn, similar to how the human brain does. Human brains are used to understanding and making viable inferences using experiences. However, in order for a machine to make an accurate prediction, the following data could be utilized. The core activity phases of a machine learning system would be - learning and inference. The discovery of patterns plays a major role. Feature selection would be the follow-up procedure, where it is decided

which of the core values of the field are put to use. The discovery part is facilitated with the collection of data, which is put to use. The right set of data is also critical at the feature selection stage. The list of these attributes is chosen by what is known as an attribute vector.The following algorithms are studied to find out which algorithm is best to implement in the system:

Supervised Machine Learning Algorithm: In Supervised learning, machine get train using data which is well "labeled." It is often compared to learning which takes place within the presence of a supervisor or an educator. A supervised learning algorithm learns from labeled training data, helps us to predict outcomes for unforeseen data. Successfully building, scaling, and deploying accurate supervised machine learning models takes time to classify the data. Moreover, Data scientist must rebuild models to form sure the insights given remains true until its data changes.

Support Vector Machine (SVM): Support Vector Machine (SVM) is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

K-Nearest Neighbor: K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. This algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. It can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available.

Artificial Neural network: A back propagation neural network was designed. In this network, 10 input neurons take as a input. The number of inputs represents the entire number of attributes in the dataset. The input layer uses Rectified Linear Unit activation function. In order to get a required recognition rate that's capable enough to diagnose the liver disorder during a patient. There is a requirement for varying certain parameters within the neural network models to produce the required optimum result.

These parameters are the learning rate, momentum rate and the hidden neurons. All these parameters present in the back propagation neural networks. The momentum rate determines the learning speed of the system.

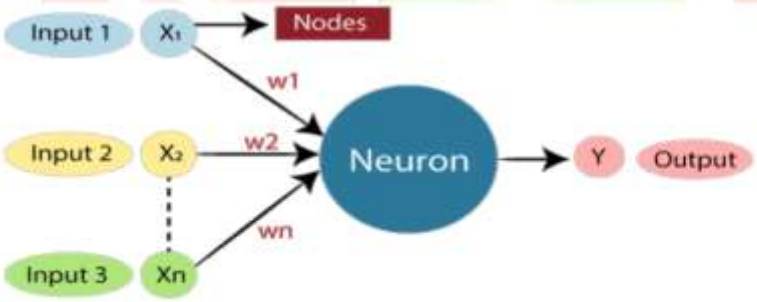

Fig-1. Artificial Neural Network

The numbers of neurons needed at the hidden layer are experimenting in order to determine the best neurons that can represent the features present within the input dataset accurately to produce the required optimum result. The numbers of neurons required within the hidden layer were experimenting by varying the neurons. The sigmoid function was utilized in the output layer due to its soft switching ability and simplicity in derivatives. The neural network was implemented in MATLAB.

IV. RESEARCH METHODOLOGY

The methodology section gives elaboration on the proposed approach. The discussion starts with the description of the overview of the approach which has been implemented over the problem cited above.The details are as follows:

Fig.1- DFD (Data Flow Diagram)

v.1 DATASET

The Indian Liver Patient Dataset comprised of 10 different attributes of 583 patients. The patients were described as either 1 or 2 on the basis of liver disease. The detailed description of the dataset is shown in Table. The table provide details about the attribute and attribute type. As clearly visible from the table, all the features except sex are real valued integers. The feature Sex is converted to numeric value (0 and 1) in the data pre-processing step.

TABLE-2 Dataset Description

No.	ATTRIBUTE	ATTRIBUTE TYPE
1.	Age	Numeric
2.	Sex	Numeric
3.	Total Bilirubin	Numeric
4.	Direct Bilirubin	Numeric
5.	Alkaline Phosphatase	Numeric
6.	Alamine Phosphatase	Numeric
7.	Total Proteins	Numeric
8.	Albumin	Numeric
9.	Albumin and Globulin Ratio	Numeric
10.	Result	Numeric[1,2]

v.2 DATA-PREPROCESSING

Data pre-processing is a crucial step of solving every machine learning problem. Most of the datasets used with Machine Learning problems need to be processed / cleaned / transformed so that a Machine Learning algorithm can be trained on it. Most commonly used pre-processing techniques are very few like missing value imputation, encoding categorical variables, normalization,standardization, etc. These techniques are easy to understand. But when we see the actual data it can be very messy. Every dataset is different and poses unique challenges. All features, except Gender are real valued integers. The last column, Disease, is the label (with '1' representing presence of disease and '0' representing absence of disease). There are in total 583 patients, where 416 are liver patient records and 167 are non-liver patient records. There contains Null value for some of the columns like Albumin and globulin ratio. The null value will be replaced with the mean of that column.

V. RESULTS AND DISCUSSION

All algorithm are similitude based on their Accuracy, Precision, Sensitivity, Specificity. After comparing we invent that artificial neural network(ANN) estimate better result than other. The patient data was reinless within the sense that the number of affected liver patients and thus the quantity of healthy individuals weren't equal. This was a difficulty during the training period. To overcome this, under-sampling and over-sampling was done. Under-sampling meant that the bulk class which during this case was the detrimental liver set was reduced to a smaller size. Over-sampling was a way during which the minority class, during this case, the good health individuals were replicated several times and combined with majority class.

Precision, Specificity and Accuracy are all measures of relevance for a binary classifier. For a binary classifier that classifies instances into positive (1) and negative (0) instances, any single prediction can fall under one among four buckets:

True Positive (TP): truth label of the given instance is positive, and thus the classifier also predicts it as a positive.
True Negative (TN): truth label is negative, and thus the classifier also predicts a negative.
False Positive (FP): truth label is negative, but the classifier incorrectly predicts it as positive.
False Negative (FN): truth label is positive, but the classifier incorrectly predicts it as negative.

Given a test set that you simply are evaluating the performance of your classifier over, our predictions will contains a bunch of TPs, TNs, FPs and FNs. Given these predictions, here is how Precision, Specificity and Accuracy are defined:

Precision: Precision is that the positive predictive value or the fraction of the positive predictions that are literally positive.

$$Precision = \frac{TP}{TP+FP}$$

Specificity: Specificity is that the true negative rate or the proportion of negatives that are correctly identified.

$$Specificity = \frac{TN}{FP+TN}$$

Accuracy: Accuracy is just the fraction of the entire sample that's correctly identified.

$$Accuracy = TP+TN$$

$$\frac{}{TP+TN+FP+FN}$$

TABLE-2 Results of classification algorithms

Classification Algorithm	Accuracy	Precision	Sensitivity	Specificity
Logistic Regression	75.67	79.57	89	35.62
K-NN	78.89	82.98	89.78	46.04
SVM	80.78	79.09	79.90	51.11
ANN	94.89	96.78	98.23	85.29

From above table, we conclude that Artificial neural network gave best result.

VI. ANALYSIS OF FACTOR AFFECTING ACCURACY

There are a set of issues that continue to challenge the accuracy of the machine learning algorithms that are used for predictive analysis.

(a)The Quantity of data that is involved:
The concise, yet precise nature of this argument being - the more the data, the more accurate the result of the predictive

analysis. With lesser data, the accuracy and effectiveness of the predictive process decline.

(b)Scope of the issue:
With machine learning paradigms demanding a huge collection of data for analysis, it is important to give due importance to the selectiveness of the features, that would pivotally define the boundaries of context, in any given problem.

(C)Parameters that are involved as a part of method: The study and analysis of the algorithm and the larger system, as a whole should also be feasible to be executed by non- technicians and absolute rookies with the basic understanding of the functioning of the system. In modern machine learning algorithms, the sense of innovation is reassured with the involvement of more than a single parameter that is involved in the analysis of the scope.

(d)Features in the data: It is imperative for any machine learning algorithm or the developer/data analyst to be able to sparsely collate the raw data and project the potentiality in the rich feature space. This is expected to accelerate the learning process of a machine learning system.
(e)Quality of Data: Any data that is to serve as a template for critical studies, fabrication, analysis and research of any subject - needs to be thoroughly checked on qualitative grounds. This is because even the slightest sense of lethargy can vandalize the integrity of the process and compromise on the potential and the expectancy, to be able to deliver.

VII. DEVELOPMENT OF GUI

Graphical user interface (GUI)) visual way of interacting with a computer. A GUI extant items that deliver information, and signify actions which can be taken by the user. GUI is created using MATLAB. MATLAB apps are independent. MATLAB programs with GUI front ends that automate a task or calculation. Many MATLAB products connect apps with consuetude user interfaces. The GUI contains input fields for all attributes within the dataset. The system will predict whether the patient has disease or not supported the trained model**.**

Fig.2- GUI development using MATLAB

CONCLUSION

In this project we classify the patient from healthy Individual. For classification we use different algorithm and implemented it in MATLAB and at the end its find out that Artificial neural network (ANN) predict high accuracy than other machine learning algorithm. The system predicts the result by using the parameter which helpful to find out the patient has liver disease or not. GUI is developed on MATLAB. MATLAB contain built in functionality which help to create GUI.

REFERENCES

[1] Joel Jacob, Johns Mathew. April-2018. Diagnosis of Liver Disease Using Machine Learning Techniques. International Research Journal of Engineering and Technology, Vol.5 Issue 4.

[2] Michael J Sorich. 2009. An intelligent model for liver disease diagnosis. Artificial Intelligence in Medicine. 47:53—62.

[3] BendiVenkataRamana, Prof. M.Surendra Prasad Babu and Prof. N. B. Venkateswarlu,. 2012. A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis". International Journal of Engineering Research and Development.