Diagnosis of Liver Disease Using Machine Learning Techniques

Joel Jacob[1], Joseph Chakkalakal Mathew[2], Johns Mathew[3], Elizabeth Issac[4]

[1,2,3] *Dept. of Computer Science and Engineering, MACE, Kerala, India*
[4] *Assistant Professor, Dept. of Computer Science and Engineering, MACE, Kerala, India*

---***---

Abstract - *Diagnosis of liver disease at a preliminary stage is important for better treatment. It is a very challenging task for medical researchers to predict the disease in the early stages owing to subtle symptoms. Often the symptoms become apparent when it is too late. To overcome this issue, this project aims to improve liver disease diagnosis using machine learning approaches. The main objective of this research is to use classification algorithms to identify the liver patients from healthy individuals. This project also aims to compare the classification algorithms based on their performance factors. To serve the medicinal community for the diagnosis of liver disease among patients, a graphical user interface will be developed using python. The GUI can be readily utilized by doctors and medical practitioners as a screening tool for the liver disease.*

Key Words: **Machine Learning, Liver Patients, Classification algorithms**

1. INTRODUCTION

Problems with liver patients are not easily discovered in an early stage as it will be functioning normally even when it is partially damaged. An early diagnosis of liver problems will increase patient's survival rate. Liver failures are at high rate of risk among Indians. It is expected that by 2025 India may become the World Capital for Liver Diseases. The widespread occurrence of liver infection in India is contributed due to deskbound lifestyle, increased alcohol consumption and smoking. There are about 100 types of liver infections. Therefore, developing a machine that will enhance in the diagnosis of the disease will be of a great advantage in the medical field. These systems will help the physicians in making accurate decisions on patients and also with the help of Automatic classification tools for liver diseases (probably mobile enabled or web enabled), one can reduce the patient queue at the liver experts such as endocrinologists.

Classification techniques are much popular in medical diagnosis and predicting diseases. Michael J Sorich [1] reported that SVM classifier produces best predictive performance for the chemical datasets. Lung-Cheng Huang reported that Naïve Bayesian classifier produces high performance than SVM and C 4.5 for the CDC Chronic fatigue syndrome dataset. Paul R Harper [2] reported that there is not necessary a single best classification tool but instead the best performing algorithm will depend on the features of the dataset to be analyzed.

The main objective of this research is to use classification algorithms to identify the liver patients from healthy individuals. In this study, FOUR classification algorithms Logistic Regression, Support Vector Machines (SVM), K Nearest Neighbor (KNN) and artificial neural networks (ANN) have been considered for comparing their performance based on the liver patient data. Further, the model with the highest accuracy is implemented as a user friendly Graphical User Interface (GUI) using Tkinter package in python. The GUI can be readily utilized by doctors and medical practitioners as a screening tool for liver disease.

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.

2. RELATED WORKS

In recent research works, several neural network models have been developed to aid in diagnosis of liver diseases in the medical field by the physicians such as diagnosis support system [3], expert system, intelligent diagnosis system, and hybrid intelligent system. In addition, Christopher N. [4] proposed a system to diagnose medical diseases considering 6 benchmarks which are liver disorder, heart diseases, diabetes, breast cancer, hepatitis and lymph. The authors developed two systems based on WSO and C4.5, an accuracy of 64.60% with 19 rules of liver disorder dataset and 62.89% with 43rules which was obtained from the WSO and C4.5respectively. Ramana [5] also made acritical study on liver diseases diagnosis by evaluating some selected classification algorithms such as naïve Bayes classifier, C4.5, backpropagation neural network, K-NN and support vector. The authors obtained 51.59% accuracy on Naïve Bayes classifier, 55.94% on C4.5 algorithm, 66.66% on BPNN, 62.6% on KNN and 62.6% accuracy on support vector machine.

The poor performance in the training and testing of the liver disorder dataset as resulted from an insufficient in the dataset. Therefore, Sug [6], suggested a method based on oversampling in minor classes in order to compensate for the insufficiency of data effectively. The author considered two algorithms of decision tree for the research work. These algorithms are C4.5 and CART [7] and the dataset of BUPA liver disorder was also considered for the experiments.

These previously designed systems have been adequate but more works has to be done on their recognition rate for better accuracy in the diagnosis of the liver disease. In this case, this will make the diagnoses of the liver diseases to be more effective and efficient by preventing misdiagnosis of the liver disorder. Developing a system with better performance than the previous works will help in preventing misdiagnosis of the disease and help in providing the best and required medication for the patient.

3. IMPLEMENTATION

i. 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-1 Dataset Description

No.	ATTRIBUTES	ATTRIBUTE TYPE
1.	Age	Numeric
2.	Sex	Nominal
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)

ii. DATA-PREPROCESSING

Data pre-processing is an important 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, scaling, etc. These techniques are easy to understand. But when we actually deal with the data, things often get clunky. 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 '2' representing absence of disease). Total number of data points is 583, with 416 liver patient records and 167 non-liver patient records. In the description of this dataset, it is observed that some values are Null for the Albumin and Globulin Ratio column. The columns which contain null values are replaced with mean values of the column.

iii. CLASSIFICATION TECHNIQUES

a) SVM

SVM aims to find an optimal hyperplane that separates the data into different classes. The scikit-learn package in python is used for implementing SVM. The pre-processed data is split into test data and training set which is of 25% and 75% of the total dataset respectively. A support vector machine constructs a hyper plane or set of hyper planes in a high- or infinite-dimensional space. A good separation is achieved by the hyper plane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

b) LOGISTIC REGRESSION

Logistic regression is one of the simpler classification models. Because of its parametric nature it can to some extent be interpreted by looking at the parameters making it useful when experimenters want to look at relationships between variables. A parametric model can be described entirely by a vector of parameters = (0, 1... p). An example of a parametric model would be a straight-line y = kx + m where the parameters are k and m. With known parameters the entire model can be recreated. Logistic regression is a parametric model where the parameters are coefficients to the predictor variables written as 0 +1 +X_1 + ...PX_p Where 0 is called the intercept. For convenience we instead write the above sum of the parameterized predictor variables in vector form as X. The name logistic regression is a bit unfortunate since a regression model is usually used to find a continuous response variable, whereas in classification the response variable is discrete. The term can be motivated by the fact that we in logistic regression found the probability of the response variable belonging to a certain class, and this probability is continuous.

c) K-NN

This section describes the implementation details of KNN algorithm. The model for KNN is the entire training dataset. When a prediction is required for a unseen data instance, the KNN algorithm will search through the training dataset for the k-most similar instances. The prediction attribute of the most similar instances is summarized and returned as the prediction for the unseen instance.

The similarity measure is dependent on the type of data. For real-valued data, the Euclidean distance can be used. Other types of data such as categorical or binary data, Hamming distance can be used. The KNN algorithm is belongs to the family of instance-based, competitive learning and lazy learning algorithms. Instance-based algorithms are those algorithms that model the problem using data in-stances (or rows) in order to make predictive decisions. The KNN algorithm is an extreme form of instance-based methods because all training observations are retained as part of the model. It is a competitive learning algorithm, because it internally uses competition between model elements (data instances) in order to make a predictive decision. The objective similarity measure between data instances causes each data instance to compete to win or be most similar to a given unseen data instance and contribute to a prediction.

d) Artificial Neural Network

A back propagation neural network was designed. In this network, 10 input neurons were present at the input layer. The number of inputs represents the total number of attributes in the dataset. The input layer uses Rectified Linear Unit activation function. The output layer contains a single layer which uses the sigmoid activation function.

In order to obtain a required recognition rate that is capable enough to diagnose the liver disorder in a patient. There is a need for varying certain parameters in 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 backpropagation neural networks. The learning rate is the learning power of the system, the momentum rate determines the learning speed of the system. The number of hidden neurons in the network has to be varied to produce the optimal result.

The numbers of neurons needed at the hidden layer are experimenting in order to deter-mine the best neurons that can represent the features present in the input dataset accurately to produce the required optimum result. The numbers of neurons required in the hidden layer were experimenting by varying the neurons. The sigmoid function was used in the output layer because of its soft switching ability and simplicity in derivatives.

The neural network was implemented using keras package which runs using the tensor-flow backend in python.

Description of the back propagation neural network is given in the table below

Table-2 ANN Description

No. of Inputs	10
No. of hidden Layers	2
No. of neurons in 1st hidden layer	400
No. of neurons in 2nd hidden layer	400
No. of Output	1
Learning Rate	0.26
Epoch	100

4. RESULTS AND EVALUATION

Our main goal going into this project was to predict liver disease using various machine learning techniques. We predicted using Support Vector Machine (SVM), Logistic Regression, K-Nearest Neighbor (K-NN) and Neural Network. All of them predicted with better results. With Each algorithm, we have observed Accuracy, Precision, Sensitivity and Specificity which can be defined as follows:

Accuracy: The accuracy of a classifier is the percentage of the test set tuples that are correctly classified by the classifier.

$$Accuracy = \frac{no.of\ TP + no.of\ TN}{no.of\ TP + FP + FN + TN}$$

Sensitivity: Sensitivity is also referred as True positive rate i.e. the proportion of positive tuples that are correctly identified.

$$Sensitivity = \frac{no.of\ TP}{no.of\ TP + no.of\ FN}$$

Precision: precision is defined as the proportion of the true positives against all the positive results (both true positives and false positives)

$$Precision = \frac{no.of\ TP}{no.of\ TP + FP}$$

Specificity: Specificity is the True negative rate that is the proportion of negative tuples that are correctly identified

$$Specificity = \frac{no.of\ TN}{no.of\ TN + FP}$$

The results of each of the classification algorithm is summarized in the table shown below.

TABLE-3 Results of classification algorithms

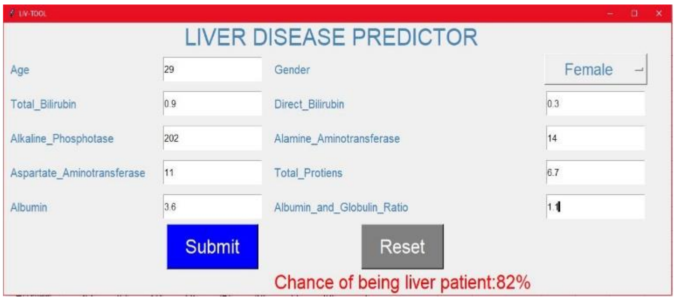
Classification Algorithm	Accuracy	Precision	Sensitivity	Specificity
Logistic Regression	73.23	78.57	88	30.62
K-NN	72.05	80.98	83.78	44.04
SVM	75.04	77.09	79	71.11
ANN	92.8	93.78	97.23	83

As clearly summarized in the table, Artificial Neural Networks gave the best results.

5. DEVELOPMENT OF GUI

The model that gave the maximum accuracy for the test data was the artificial neural network. So, Artificial neural network is used for creating the GUI. The GUI is created using Tkinter package in python. Two GUIs are created, one for predicting and the other for training new data. The GUI contains input fields for all attributes in the dataset. The system will predict whether the patient has liver disease or not based on the trained model. The GUI will be a useful tool for medical staff in the early diagnosis of liver disease in patients. A picture of the developed GUI is shown below.

Fig-1 GUI developed using python



LIVER DISEASE PREDICTOR

Age	29	Gender	Female
Total_Bilirubin	0.9	Direct_Bilirubin	0.3
Alkaline_Phosphotase	202	Alamine_Aminotransferase	14
Aspartate_Aminotransferase	11	Total_Protiens	6.7
Albumin	3.6	Albumin_and_Globulin_Ratio	1.1

Chance of being liver patient:82%

6. CONCLUSION

In this project, we have proposed methods for diagnosing liver disease in patients using machine learning techniques. The four machine learning techniques that were used include SVM, Logistic Regression, KNN and Artificial Neural Network. The system was implemented using all the models and their performance was evaluated. Performance evaluation was based on certain performance metrics. ANN was the model that resulted in the highest accuracy with an accuracy of 98%. Comparing this work with the previous research works, it was discovered that ANN proved highly efficient. A GUI, which can be used as a medical tool by hospitals and medical staff was implemented using ANN.

REFERENCES

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

[2] Paul R. Harper, A review and comparison of classification algorithms for medical decision making.

[3] BUPA Liver Disorder Dataset. UCI repository machine learning databases.

[4] Prof Christopher N. New Automatic Diagnosis of Liver Status Using Bayesian Classification.

[5] Schiff's Diseases of the Liver, 10th Edition Copyright ©2007 Ramana, Eugene R.; Sorrell, Michael Maddrey, Willis C.

[6] P. Sug, On the optimality of the simple Bayesian classifier under zero-one loss, Machine Learning 29 (2–3) (1997) 103–130.

[7] 16th Edition HARRISON'S PRINCIPLES of Internal Medicine