International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 07 | July 2021 www.irjet.net p-ISSN: 2395-0072

Detection of Liver Diseases Using Classification Models

Sunidhi S Parvatikar[1]

[1]*Basaveshwara Engineering College, Vidyagiri, Bagalkot, Karnataka, India*

Abstract: *Liver diseases result approximately 2 million deaths per year worldwide. A Liver infection causing Chronic Hepatitis B affects around 257 million people around the globe as reported by WHO. Hence, liver disease is one the most dangerous disease and is required to be treated as early as possible. Machine learning methodologies have been advanced on various liver disease related datasets to expect result. In this application, we have used more than 500 Indian patients' dataset as input. The models which are built on top of lower dimensional data are more reasonable and explainable. Thus, experts in machine learning today can guarantee an exact and definite diagnosis and analysis of a disease. People can't simply detect the liver disease; hence I propose a classification model to detect the liver diseases by objective features of the data. Consequently, results obtained were quite satisfactory with 92.95% of accuracy using regression model.*

Keywords: Artificial Intelligence, Liver disease, Machine Learning, Classification algorithms.

1.INTRODUCTION

Liver problems are caused by a various factor that harm the liver, such as viruses, substance abuse and obesity. Liver disease doesn't always cause noticeable signs and symptoms. If the signs and symptoms of liver disease do occur, [2] they may include: the changing color of skin and eyes to yellow(jaundice), Abdominal pain and swelling in the legs and ankles, itchy skin, dark urine color, pale stool color, chronic fatigue, nausea or vomiting, loss of appetite, tendency to bruise easily and many more. During the early stages of liver illness, it is exceptionally hard to identify even though liver tissue has already been harmed. It requires numerous specialists to analyze the damage [4]. Hence, early diagnosis of liver problems will increase patient's survival rate.

There are 4 liver disorder stages:

- Fatty liver is a painful liver condition portrayed by liver irritation and arrangement of scar tissue, which has numerous conceivable causes, including corpulence, poor nourishment and consumption of medication without consultation, among numerous others. It can happen in individuals with an abnormal state of liquor consumption as well as in people who never had liquor [3].
- Cirrhosis is another important type of liver damage. It is usually the result of long-term damage of liver. When liver is damaged for a long time and starts to malfunction this particular type of liver damage occurs. Each time your liver is injured — whether by disease, excessive alcohol consumption or any other cause — it tries to repair itself. In the process, scar tissue forms. As cirrhosis progresses, more and more scar tissue form, making it difficult for the liver to function [2]. This may result fatal.
- Hepatitis is usually caused by an infection that spreads by direct contact with tainted body [3]. Today, chronic HCV is usually curable with oral medications taken every day for two to six months [2].
- Liver Cancer risk is higher on those who has cirrhosis. Most often it spreads from liver to other organs [3].

There are more than 100 types of liver infections. Therefore, developing a machine that will specialize in the diagnosis of the disease will be of a boundless advantage in the medical field [5].

As cited above, it is extremely hard to notice the liver disease and requires a lot of experts. Thus, an identification of liver disease is not at all easy, accordingly an effort has been made to perceive these liver infections with the use of machine learning techniques. There are many classification techniques in machine learning used for detecting or predicting diseases in the medical field. We can use any classification algorithms such as Logistic Regression, SVM (Support Vector Machine), Naïve Bayes, decision trees, random trees and many more. There is no necessity of a single best classification tool but instead the best performing algorithm will depend on the features of the dataset to be analyzed as reported by Paul R Harper [5].

The main objective of this research, by using the classification algorithms to detect if the patient is suffering from the liver disease from dataset of individuals by considering the features of the liver test such as total bilirubin, direct bilirubin, alkaline phosphatase alanine aminotransferase, aspartate aminotransferase, total proteins, albumin, albumin and globulin ratio, gender and age. Then by using five classification algorithms such as Support Vector Machines (SVM), logistic regression, decision trees, Artificial Neural Networks (ANN) and K-nearest

neighbor (KNN) we implement on datasets, further we select the highest accuracy performing model.

2. RELATED WORKS

In recent research works, several classification algorithms models have been developed to aid in diagnosis of liver diseases in the medical field by the physicians [5] such as diagnosis support system, expert system, intelligent diagnosis system, and hybrid intelligent system.

The authors [5] made a study on liver diseases diagnosis by examining some classification algorithms such as Support Vector Machine (SVM), Logistic Regression, K-Nearest Neighbor (K-NN) and Neural Network. The authors obtained the accuracy of 73.23% on logistic regression, 72.05% on K-NN, 75.04% on SVM, 92.8% on ANN.

Also, the authors [18] completed a serious study on liver diseases analysis by estimating 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. These formerly designed machines are acceptable but more works has to be done on their acknowledgment rate for better accuracy in the diagnosis of the liver disease.

Developing a machine with better performance than the previous works will aid in preventing misdiagnosis of the disease and help in providing the best and required medication for the patient.

3. IMPLEMENTATION

In this research, I have used dataset of about 500+ Indian patients having about 11 features, making the dataset size (583, 11). Below (fig.1) is the block diagram of the approach followed in this research.

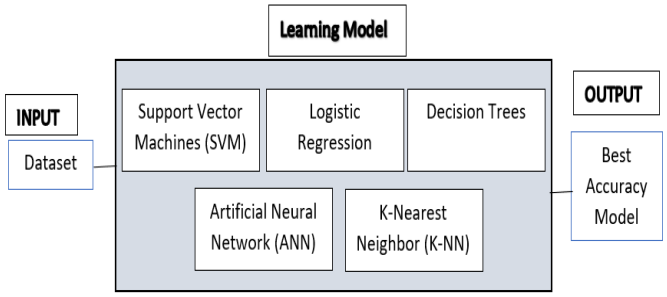

Fig.1. Block diagram of the approach.

3.1. Dataset:

The dataset which is used as input, can be viewed as a collection of data objects, which are often also called as records, points, vectors, patterns, events, cases, samples, observations, or entities. Data objects are described by a number of features, that capture the basic characteristics of an object, such as the mass of a physical object or the time at which an event had occurred and many more [13].

The dataset has both liver disease and non-liver disease individuals which was collected form North-East Andhra Pradesh, India. The objective is to recognize the liver disease patients from the dataset by training the machine.

3.2. Data-preprocessing:

Primarily, I conducted a dataset preprocessing. Data preprocessing is a data mining technique that involves transforming raw data into an understandable format [8]. We will pre-process the data so that we can use it in our code efficiently [12]. Every dataset is different and poses unique challenges. Table.1 shows the features of the dataset used in liver detection. Last row, Dataset (Result), is the label with '1' representing presence of disease and '0' representing absence of disease.

As we know, machines do not comprehend free text, image or video data as it is, they understand 1s and 0s. So, it probably would not be good enough if we just put on a slideshow of all our images and expect our machine learning model to get trained just by it [13].

Table - 1: Dataset of Indian patients.

No.	Features	Features type
1.	Age	Numeric
2.	Gender	Nominal
3.	Total bilirubin	Numeric
4.	Direct bilirubin	Numeric
5.	Alkaline phosphatase	Numeric
6.	Alanine aminotransferase	Numeric
7.	Aspartate aminotransferase	Numeric
8.	Total proteins	Numeric
9.	Albumin	Numeric
10.	Albumin and globulin ratio	Numeric
11.	Dataset (Result)	Numeric (0,1)

The features of the data can be easily interpreted by the algorithm when we pre-process the data.

During the data preprocessing, initially load the dataset by importing the required libraries. Dataset used is a CSV file, because they are of lightweight. Then we eradicate the missing values from the dataset, if not handled properly we might end up getting inaccurate results. After removing the duplicates and null values, we then visualize, below (fig.2) is a bar plot, based on the frequency of features in the dataset. This bar plot of frequency of features would be a lot helpful to train the machine.

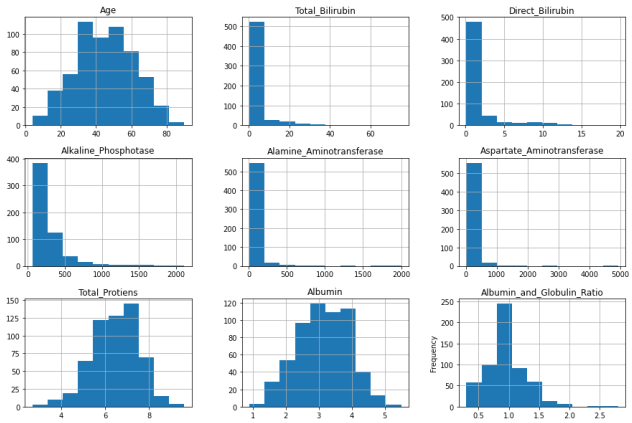

Fig.2. Dataset features with frequency

Then we have to perform the feature sampling which is a very common method for selecting a subset of the dataset that we are analyzing. In most cases, working with the complete dataset can turn out to be too expensive considering the memory and time constraints. Using a sampling algorithm can help us reduce the size of the dataset to a point where we can use a better, but more exclusive machine learning algorithm [13].

Later, after feature encoding is also done, split the data into testing and training samples. In this research, dataset is divided into 70:30 ratio on each algorithm, where 70% of data is used for training and 30% of data is used for testing. The algorithm model is going to learn from the data and then the make required predictions.

3.3. Classification techniques:

3.3.1. Support Vector Machine (SVM):

SVM is mostly used for classification problems. SVM tries to find the best decision boundary in such a way that reduces the risk of errors on data [6]. A simple SVM diagram with optimal hyperplane with two classes of data is as shown (fig.3) that maximizes the margin between the two classes [14]. SVM uses kernel to find the decision boundary. In the SVM algorithm, we are looking to maximize the margin between the data points and the hyperplane.

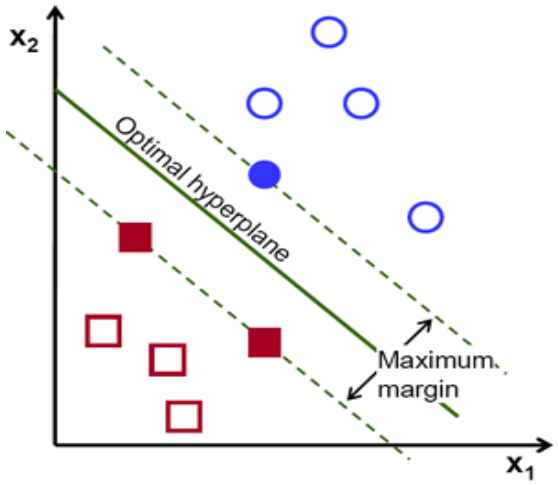

Fig.3. Support Vector Machine with two data classes.

3.3.2. Logistic Regression:

Logistic regression is a type of a supervised machine learning algorithm. It makes a prediction that has binary outcome from the past data. Logistic regression is applied to an input variable (X) where the output variable (y) is a discrete value which ranges between 1 (yes) and 0 (no) [6]. It is a predictive analysis algorithm and based on the concept of probability. The hypothesis of logistic regression tends it to limit the cost function between 0 and 1 [15]. i.e.,

$$0 \leq h_\theta(x) \leq 1$$

In order to map predicted values to probabilities, we use the Sigmoid function. The function maps any real value into another value between 0 and 1. In machine learning, we use sigmoid to map predictions to probabilities [15]. Below (fig.4) shows the sigmoid function for logistic regression.

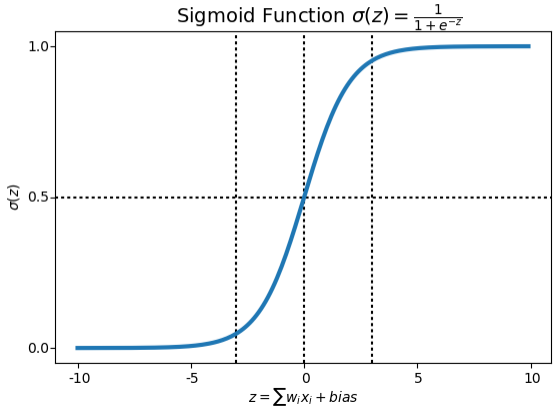

Fig.4. Sigmoid function graph.

3.3.3. Decision tree:

Decision tree is a type of supervised learning model which splits the dataset into two or more sets, which includes two common parameters namely nodes and leaves. The result of this algorithm is either "true" or "false".

In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node [16]. For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree [16].

3.3.4. Artificial Neural Network (ANN):

ANN is foundation of artificial intelligence (AI) and it initially goes through a training phase where it learns to recognize patterns in data, whether visually, aurally or textually. During this supervised phase, the network compares its actual output fashioned with the desired output [7]. A simple neural network can be represented as shown in the figure below:

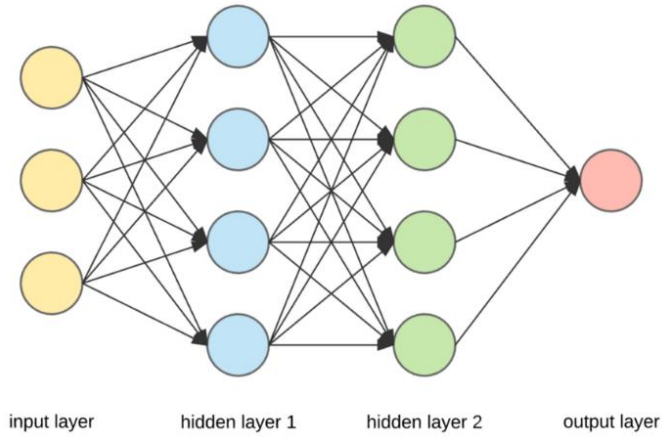

Fig.5. Simple Neural Network

3.3.5. K- Nearest Neighbor (KNN):

K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique. This algorithm accepts the similarity between the new case/data and available cases and put the new case into the group that is most similar to the available categories [12]. This algorithm supplies all the available data and classifies a new data point based on the similarity. This means when new data appears then it can easily be classified into a well-suited category by using KNN algorithm.

KNN algorithm can be used for regression as well as for classification but mostly it is used for the classification problems [12].

3.3.6. Best accuracy model:

After applying the train and test on each algorithm model, we have to check which algorithm model is more suitable by performance metrics, so that our machine can make decent predictions. Then we have to select a good predicting model based on the accuracy of the algorithm therefore we can have the finest and more suitable one (Table.2).

Data analysis algorithms work better if the dimensionality of the dataset is lower. As the models which are built on top of lower dimensional data are more reasonable and explainable. The data may now also get easier to visualize (fig.6). Features can always be taken in pairs or triplets for visualization purposes, which makes more sense if the feature set is not that big [13]. Thereafter selecting the model, the frequency of people suffering from liver disease and people who aren't suffering from the liver disease can be recognized.

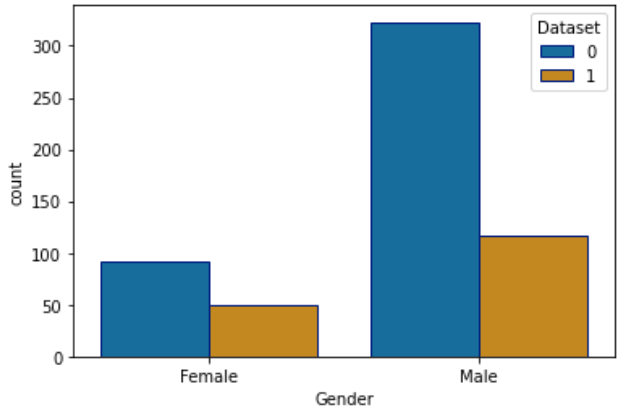

Fig.6. Frequency of male and female individuals with or without liver disease.

The bar graph (fig.6) visualizes the number of male and female individuals from the dataset with or without liver disease (with '1' representing presence of disease and '0' representing absence of disease). Thus, the machine now can classify the individuals from the given dataset by producing as good classification model.

Based on the classification algorithms accuracies, the logistic regression model has to be nominated, as it has resulted the highest accuracy and even the accuracy is same as the precision. Likewise, it is one of the most popular Machine Learning algorithms.

Therefore, after examining different classification algorithms, we have concluded with logistic regression,

which produces a good classification model. Thus, this machine will be a great application in the medical field which can detect patients who can even be in early stage and help them to increase their survival rate.

4. RESULT AND DISCUSSIONS

The main objective in this research was to predict liver disease individuals using various machine learning techniques. I have predicted using Support Vector Machine (SVM), Logistic Regression, Decision trees, K-Nearest Neighbor (KNN) and Artificial Neural Network (ANN). All of them predicted with better results (Table.2) As clearly summarized in the table logistic regression gave the best results.

Table-2: Performance of classification algorithms

Classification Techniques	Accuracy	Precision
Support Vector Machine (SVM)	72.87%	76.01%
Logistic Regression	92.95%	93.59%
Decision Trees	82.73%	85.20%
Artificial Neural Networks (ANN)	81.78%	83.82%
K-Nearest Neighbor (K-NN)	85.09%	87.15%

Finally, (fig.7) I can conclude that logistic regression algorithm has managed to produce finest classification model with 171 liver disease patients from 583 total number of individuals. Comparing this work with the previous research works, it was discovered that logistic regression proved highly efficient.

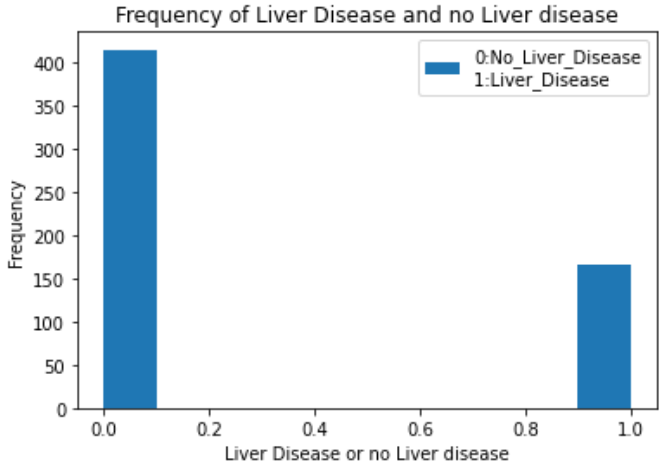

Fig.7. Total no of individuals with and without liver disease.

This machine will be an application which will bring about a dynamic change in the field of medical sciences, as it will predict the liver patients in the early stage and can reduce the life risk. Overall, 1 million individuals die due to complications of cirrhosis. Cirrhosis (liver disease) is currently the 11th most common cause of death globally [9]. As liver disease is not easily perceptible, this machine can help the individuals to get the treatment in their early stage, if they reach the last stage it can't be undone.

Our machine has a produced an accuracy which is not only satisfactory but worth applauding. Therefore, if this machine is introduced in the medical field, world-wide death count due to liver disease can be reduced. Besides cirrhosis might not be in world's top 20 most common death cause disease.

5. CONCLUSIONS

In this research, we successfully classify the liver disease patients from non-liver disease patients with 92.95% of accuracy. The five machine learning techniques that were used include SVM, Logistic Regression, decision trees, Artificial Neural Network (ANN) and K-Nearest Neighbor (KNN). The system was implemented using these models and their performance was evaluated. Performance evaluation was based on certain performance metrics. All the models produced good accuracy for the given dataset. And the highest accuracy model which is logistic regression was selected as finest classification model.

In today's world, Artificial Intelligence applications in medical field is a boon. Implementation of this model in the medical field can benefit many individuals and can also reduce the hazard.

Why do we invest billions of dollars every year in technology if it does not save lives? It's a question worth asking. We all have to ponder on this because there is no greater cause than saving a human life. We have realized this in these times of pandemic where we have learned the value of every human life around us.

ACKNOWLEDGEMENT

The author would like to express gratitude for guiding to prepare this research paper provided by Prof V. B. Pagi, The Head of Department of Computer Science, Basaveshwara Engineering College.

REFERENCES

[1] Anatomy and function of the liver at medicinenet by Medical Author: Benjamin Wedro, MD, FACEP, FAAEM, Medical Editor: Bhupinder S. Anand, MBBS, MD, DPHIL (OXON) and Medical Editor: Melissa Conrad Stöppler, MD.

[2] Mayoclinic, liver disease, Mayo foundation for Medical Education and Research.
[3] Machine Learning Techniques on Liver Disease - A Survey V.V. Ramalingam, A. Pandian, R. Ragavendran, International Journal of Engineering & Technology.
[4] Healthline, Stages of liver failure, medically reviewed by Saurabh Sethi, M.D., MPH — Written by Jill Seladi-Schulman, Ph.D. on April 9, 2019.
[5] Joel Jacob, Joseph Chakkalakal Mathew, Johns Mathew, Elizabeth Issac Dept. of Computer Science and Engineering, MACE, Kerala, India Assistant Professor, Dept. of Computer Science and Engineering, MACE, Kerala, India. International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 05 Issue: 04 | Apr-2018.
[6] Medium, Patricia Bassey, Axum labs, Research Lab for Axum Technologies Ltd, Logistic Regression Vs Support Vector, Sept-19-2019.
[7] Investopedia, by Jake F, Reviewed by ERIC E, Artificial Neural Network, Aug 28, 2020.
[8] Hackernoon, what steps should one take while doing data-preprocessing Mohit Sharma, July 25,2018.
[9] Pub med, Burden of liver diseases in the world, Sumeet K Asrani, Harshad Devarbhavi , John Eaton, Patrick S Kamath, Sept 26, 2018.
[10] towardsdatascience, Support Vector Machines-Introduction to learning algorithms, Rohit Gandhi, June 7, 2018.
[11] towardsdatascience, Applied deep learning part:1 Artificial neural network, Arden Dertat, Aug 8,2017.
[12] Javatpoint, KNN algorithm for Machine learning [13] towardsdatascience, Data preprocessing concepts by pranjal pandey, Data Science Enthusiast · Data & Analytics Consultant, Nov 25, 2019.
[13] D. Xu, H. Fu and W. Jiang, "Research on Liver Disease Diagnosis Based on RS_LMBP Neural Network," 2016 12th International Conference on Computational Intelligence and Security (CIS), Wuxi, 2016.
[14] towardsdatascience, Introduction to logistic regression by Ayush Pant, Jan 22,2019.
[15] javatpoint, Decision tree classification algorithm.
[16] P. Sug, On the optimality of the simple Bayesian classifier under zero-one loss, Machine Learning 29 (2–3) (1997) 103–130.
[17] Schiff's Diseases of the Liver, 10th Edition Copyright ©2007 Ramana, Eugene R.; Sorrell, Michael Maddrey, Willis C