IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 6, June 2022

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.118

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106019 |

Liver Disease Prediction Using Machine Learning

P. Radha Vyshnavi[1], M.V.N. Sai Niharika[2], M. Summayya[3], P. Pravallika[4], Dr. Sri Hari Nallamala [5]

U.G. Student, Department of Computer Science and Engineering, Vasireddy Venkatadri Institute of Technology, Namburu, Andhra Pradesh, India [1,2,3,4]

Associate Professor, Department of Computer Science and Engineering, Vasireddy Venkatadri Institute of Technology, Namburu, Andhra Pradesh, India [5]

ABSTRACT:Liver plays a significant function in metabolism. It is in danger of several diseases like liver cancer, chronic hepatitis and cirrhosis. It is very difficult to effectively predict liver disease using large amounts of medical data. Early diagnosis, and treating the patients are compulsory to reduce the risk of those deadly disease.

Numerous researches have been performed using Machine Learning algorithms for classifying liver disorder cases. The main aim of this research work is to predict liver disease using different Machine Learning algorithms such as K-Nearest Neighbor(KNN), Support Vector Machines(SVM), Random Forest etc. To get more accuracy and believability we are using Ensemble model of the above mentioned machine learning algorithms.

KEYWORDS – Machine Learning, K-Nearest Neighbor, Support Vector Machines, Random Forest, Artificial Neural Networks(ANN), Ensemble model.

I. INTRODUCTION

We all know that liver is one among the largest organ in human body which performs vital body function including making proteins, blood clotting, manufacturing triglyceride and cholesterol, glycogen synthesis and bile production. Usually, over 75% of liver tissue must be stricken by a decrease in function to occur. So it is important to detect the disease at an early stage such that the diseases are often treated before it becomes severe. Liver diseases aren't easily discovered in an early stage because they appear like functioning normally even when it's partially damaged. An early diagnosis of liver diseases will increase the patient's survival rate. There are about 2 million deaths due to liver diseases each year worldwide. The widespread occurrence of disease in India is accounted because of inactive lifestyle, increased alcohol consumption and smoking. There are about 100 kinds of liver infections. Therefore, developing a machine that may enhance within the diagnosis of the disease are going to be of a good advantage within the medical field. These systems will help the physicians in making accurate decisions on patients and also with the assistance of classification tools for liver diseases, one can reduce the patient queue at the liver experts. Machine Learning Classification techniques now-a-days became greatly important within the healthcare sector for the prediction of disease from the medical database. The main objective of this research is to use classification algorithms to spot the liver patients from healthy individuals. during this study, Random Forest(RF), Support Vector Machines (SVM), K Nearest Neighbor (KNN) and Artificial Neural Networks (ANN). An Ensemble model are considered to urge more accuracy and believability and for comparing their performance supported the liver patient data. Further, the model with the very best accuracy is implemented as a user friendly Graphical interface (GUI) using Flask module in Python. The GUI are often readily employed by doctors and medical practitioners as a prediction tool for disease.

II. RELATED WORK

In recent research works, several models were developed to support prediction of liver diseases within the medical field. Hoon Jin etc, described the concept of various classification techniques that help the doctors to identify the disease quickly and efficiently. Various classifiers like Naïve Bayes, Multi-Layer Perceptron, Decision Tree and k-NN were compared and analysed based on several parameters and also the dataset was collected from UCI Repository. The experimental results showed that in terms of precision, Naïve Bayes gave the higher classification results whereas

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106019 |

Logistic Regression and Random Forest gave better ends up in terms of recall and sensitivity. Christopher N. proposed a system to diagnose medical diseases considering 6 benchmarks which are liver disorder, heart diseases, diabetes, carcinoma, hepatitis and lymph. Ramana also made a study on liver diseases diagnosis by evaluating some selected classification algorithms like Naive Bayes classifier, backpropagation neural network, K-NN and support vector. They got result as 51.59% accuracy on Naive Bayes classifier, 66.66% on BPNN, 62.6% on KNN and 62.6% accuracy on SVM. The training and testing of the liver disease dataset leads to poor performance that resulted in insufficiency within the dataset. Therefore, Sug, suggests a technique supported oversampling in minor classes so as to complete the insufficiency of information effectively. The author considered two decision tree algorithms for his research work, the algorithms are C4.5 and CART and also BUPA liver disorder dataset was considered for this experiments. The past designed systems are adequate but more works has got to be done on their recognition rate for better accuracy within the diagnosis of the disease. during this case, this can make the diagnoses of the liver diseases to be simpler and efficient by preventing mis-diagnosis of the liver disorder. To develop a system with better performance than the previous works will help in preventing misdiagnosis of the disease and help in providing the most effective and required treatment and medicine for the patients.

III. PROPOSED SOLUTION

i. DATASET

The Indian Liver Patient Dataset comprised of 11 different attributes of 583 patients i.e 10 different attributes and one class attribute. The patients were described as either 1 or 2 on the premise of liver disease. All the features are Numeric i.e is real valued integers except the Gender attribute. Gender attribute is of Nominal type it contains female and male data. The feature Gender is converted to numeric value (0 and 1) with in the data pre-processing step. Below mentioned Table describes the detailed information about the dataset. The table provide dataset information such as attributes and attribute type.

ATTRIBUTES	TYPE
Age	Numeric
Gender	Nominal
Total Bilirubin	Numeric
Direct Bilirubin	Numeric
Alkaline Phosphatase	Numeric
Alamine Phosphatase	Numeric
Aspartate Aminotransferase	Numeric
Total Proteins	Numeric
Albumin	Numeric
Albumin & Globulin Ratio	Numeric
Dataset	Numeric

ii. Data Pre-processing

Data Pre-processing is the main step to prepare the data as of model requirements. The preliminary step for any dataset is to be pre-processed before applying to any algorithm. The general steps involved in Data Pre-processing

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106019 |

are: ”Getting dataset, Importing libraries, Importing datasets, Finding Missing Data, Encoding Categorical Data, Splitting dataset into training and test set, Feature Scaling”. Our dataset involves the following steps.

A) Handling Missing Data

There are two ways for handling missing data they are: By Deleting the particular row and By Calculating the mean. In calculating mean, we calculate the mean of that column or row which contains any missing value and replace the mean value with missing value. This is useful for the features which have numeric data. In our work, we handled the missing data by filling it with median value.

B) Encoding Categorical Data

Machine learning completely work on mathematics and numbers. Suppose if our dataset has categorical variable then it may create problem while building model. Therefore,we should make sure to encode the categorical variables into numbers.
In our Dataset, Gender has Nominal type so we convert Male to 1 and Female to 0.

C) Splitting dataset into training and test set

The whole dataset is divided into training dataset and testing dataset, in which we train the algorithm with the training dataset and test the algorithm with the testing dataset. Test dataset is used later to determine the efficiency of each algorithm. In our research work, we divide the dataset with test_size=0.25 i.e 75% as train data and 25% as test data

D)Feature Scaling

This is the last step in data pre-processing in machine learning. It's a way to standardize the independent variables of the dataset in an exceedingly specific range. In this, we put our variables within the same range and within the same scale so no variable dominates the opposite variable.

IV. APPLYING ML ALGORITHMS

Researchers has worked on implementing each of the algorithms. So, the used classifiers in our work are Random Forest Classifier, Support Vector Classifier, K Nearest Neighbor Classifier, Artificial neural networks (ANN) and Weighted Averaging ensemble model is additionally designed for more believability.

A) Support Vector Classifier

Support Vector Machine(SVM) algorithm is useful for Classification and also for Regression problems. it's mainly used for Classification problems in Machine Learning. There are two main advantages of SVM one is higher speed and therefore the other is for best performance with a limited number of samples. A simple linear SVM classifier is implemented by making a line between two classes by which the points that are one side of the straight line belongs to one category and the data points on the opposite side of the straight line belongs to a different category. The aim of the SVM algorithm is to make the foremost effective line or decision boundary which will separate n-dimensional space into classes so as that we will easily put the new data point within the right category.

Algorithm:

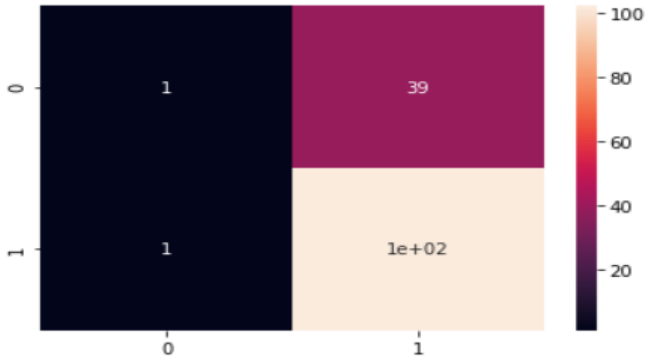
1. The SVM algorithm is imported from the scikit-learn package.
2. Split data into training and test data.
3. Generate a SVM model.
4. Train or fit the data into the model.
5. Predict the review

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106019 |

Result:



From the confusion matrix of Support Vector machine(SVM) model, accuracy has been calculated and it comes out to be 72.02%. Precision value is resulted as 0.723, F1-Score value is resulted as 0.836 and the Recall value comes out to be 0.990 .

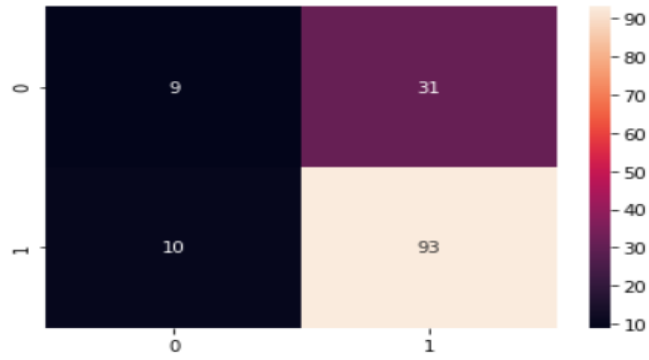
B) K Nearest Neighbours Classifier

K Nearest Neighbors classifier is the simplest and most widely used machine learning algorithm. KNN algorithm is used to solve classification and regression problems in Machine Learning. However, it is mainly used for classification problems. KNN algorithm can be implemented by finding a group of k objects which are nearest to the test object, and a label is assigned based on predominance of a class in the neighbourhood of the object.

Algorithm:

1. Import k-nearest neighbor algorithm from scikit-learn package.
2. Create feature and target variables.
3. Split the data into training and test data.
4. Generate a k-NN model.
5. Train or fit the data into the model.

Result:



From the confusion matrix of K Nearest Neighbour (KNN) model we have calculated the accuracy value and it is resulted as 71.32%. Precision value is resulted as 0.750, F1-Score value is resulted as 0.819 and Recall value comes out to be 0.902.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106019 |

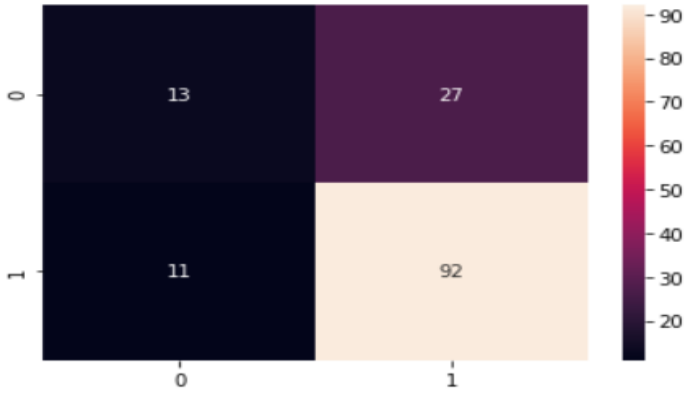
C) Random Forest Classifier

Random Forest is one of the most popular machine learning algorithm which belongs to the supervised learning technique. Random Forest Algorithm can be used for both Classification and Regression problems in Machine Learning. Random Forest classifier contains many number of decision trees of the given dataset which based on various subsets and then it takes the average to improve the accuracy of that dataset. Instead of hoping on one decision tree, the random forest takes the prediction from each tree and it predicts the final output is based on the majority votes of predictions.

Algorithm:

1. Import Random Forest algorithm from scikit-learn package.
2. Create feature and target variables.
3. Split the data into training and test data.
4. Generate a random forest model.
5. Train or fit the data into the model.

Result:



From the confusion matrix of Random Forest Classifier model, accuracy has been calculated and it comes out to be 73.42%. Precision value is resulted as 0.773, F1-Score value is resulted as 0.828 and the Recall value comes out to be 0.893.

D) Artificial Neural Networks

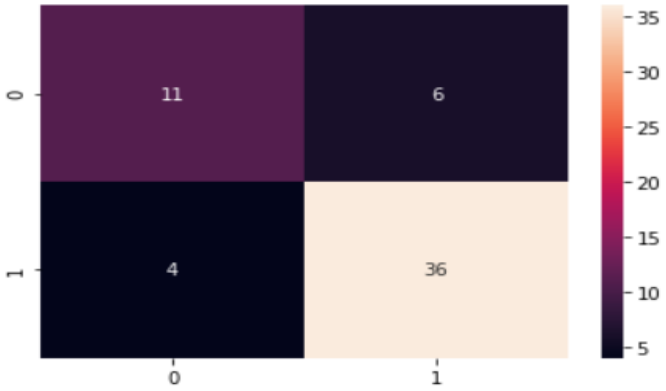
An artificial neural network (ANN) is a computational model to perform tasks like classification, decision making, prediction, etc. It consists of artificial neurons. A back propagation neural network was designed. This network consists of 10 input neurons present in the input layer. The number of inputs in the network represents the total number of attributes in the dataset. The output layer uses a sigmoid activation function which contains a single layer. So as 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 within the neural network models to produce the required optimum result. The neural network was implemented using keras package which runs using the Tensor-flow backend in python.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106019 |

Result:



From the confusion matrix Random Forest Classifier model, accuracy has been calculated and it comes out to be 82.45%. Precision value comes out to be 0.857, F1-Score value come out to be 0.878 and Recall value comes out to be 0.900

E) Ensemble Model

Ensemble models are the techniques that aim at improving the accuracy of results in models by combining multiple models rather than using a single model. The combined models increase the accuracy of the results when compared to the accuracy obtained by the single model. Ensemble models also helps in increasing the believability of the model. This advantage has boosted the popularity of ensemble methods in machine learning. Ensemble model is a solution to overcome some of the technical challenges while using a single estimator they are High variance and Low accuracy.

The basic ensemble model techniques are MaxVoting, Averaging, Weighted Average and there are some advanced ensemble model techniques like Stacking, Blending, Bagging and Boosting.

In this work, we are using Weighted Average basic ensemble model. Weighted Average ensemble model is an extension of the averaging method. Other models are assigned different weights defining the importance of each model for prediction where the prediction of every model is multiplies by the weight and then it's average is calculated. This model resulted with an accuracy of 73.33%

V. RESULT AND EVALUATION

SVM Classifier, k-Nearest Neighbours, Random Forest Classifier and Artificial Neural Networks are four types of machine learning classifiers used in this research. The operations of splitting the data set into two parts were tested for a data set which is obtained from preprocessing the dataset. The below table gives the Accuracy Score, Precision score, F1-Score and Recall Score for each classifier that is performed on the dataset.

Accuracy: Accuracy of a classifier represents the number of correctly classified data instances over the total number of data instances.

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

Precision: Precision is defined as the ratio of the true positives to that of positive results i.e is both true positives and false positives

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106019 |

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

Recall: Recall is also referred as True positive rate or Sensitivity i.e. it is the ratio of True positives to that of the tuples that are correctly identified.

$$Recall = \frac{TP}{TP + FN}$$

F1-Score: F1-score is a measure which takes both precision and recallvalues into consideration.

$$F1\text{-}score = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

	Model	Accuracy	Precision	F1	Recall
0	KNeighborsClassifier	0.713287	0.750000	0.819383	0.902913
1	SupportVectorClassifier	0.720280	0.723404	0.836066	0.990291
2	RandomForestClassifier	0.734266	0.773109	0.828829	0.893204
3	Artificial Neural Networks	0.824561	0.857143	0.878049	0.900000

VI. CONCLUSION

The research paper has an effort to reveal that the classification techniques serve as powerful tools for liver disease prediction. By applying different classification algorithms like Support Vector Machine, K Nearest Neighbour Classifier, Random Forest Classifier, and Artificial Neural Networks. The machine was implemented using all the models and their performance was measured. Performance evaluation was based on certain performance metrics. ANN was the model that resulted in the highest accuracy with an accuracy of 82%. It is proven that Support Vector Machine gave better results in terms of Recall with a Recall of 0.990. ANN model shown better results in terms of accuracy, precision and f1-score. Finally this research work helps in identifying better classifiers among Support Vector Machine, k-Nearest Neighbours, Random Forest , and Artificial Neural Networks for prediction of liver disease.

REFERENCES

[1] BendiVenkataRamana, Surendra. Prasad Babu. M, Venkateswarlu. N.B, A Critical Study on liver disease diagnosis by Selected Classification Algorithms, International Journal of Database Management Systems (IJDMS), Vol.3, No.2, May 2011

[2] Hoonjin,Seoungcheonkim, Jinhong Kim, “Decision Factors on Effective Liver Patient Data Predictions “ Science and Bio-Technology, Vol. 6, Issue.4, pp. 167-178, 2014.

[3] Ayesha Pathan, Diksha Mhaske2, ShrutikaJadhav, RupaliBhondave, Dr.K.Rajeswari, - Comparative Study to predict Liver disease using different Classification Algorithms on ILPD Dataset, International Journal for Research in Applied Science & Engineering Technology (IJRASET), Vol. 6, Issue.2, pp. 388-394, 2018.
TapasRanjanBaitharu, Subhendu Kumar Pani

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

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

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106019 |

[6]Comparative Study of Artificial Neural Network(ANN) based Classification of Liver Patient,Journal of Knowledge Engineering and Application.
[7]Reetu and N.Kumar (2015),Liver Cancer diagnosis Using Classification Techniques, International Journal of Recent Scientific.Volume 6. Issue 6.
[8]L.A.Auxilia,"Prediction of Accuracy using machine learning techniques for Indian patient liver diseases," in 2 nd ICOEI, Tirunelveli, India, pp. 45–50, 2018.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details