A Comparative Analysis of Classification Algorithms in Liver Disease Detection

Akanksha Soni^{1*}, Anivash Rai²

^{1,2}Department of Electronics and Communication Engineering,
UIT, RGPV, Bhopal, Madhya Pradesh, India

akanksha.sonivits@gmail.com,

avinashrai@rgtu.net

Abstract

Liver is the prime organ in the human body and executes a number of functions related to producing necessary proteins, carbohydrates metabolism, energy storage, and detoxification of waste. Diseases that may affect the liver include hepatitis, fatty liver, bleeding, fatigue, jaundice. Liver disease affects millions of the populace around the earth and diagnosis at an initial phase is very essential for better treatment. It is an extremely difficult task for medical researchers to predict the infection in the prior stages owing to subtle symptoms. So there is a solid requirement for an efficient, accurate, and practical framework to predict the consequence of such infection. It will be valuable for taking precautions and appropriate treatment. To reduce this concern, we proposed a system that is capable enough to detect liver disease using machine learning approaches. The major intention of this work is to use classification algorithms to identify the liver patients from healthy individuals with the help of automated processes to overcome human fault and decrease the false alarm rate that gives the progression which is more accurate, effortless, and unproblematic. This project also aims to compare the categorization algorithms that depend on their performance features. To gain results with less processing instance and a high precision rate. In this work, four classification algorithms that are Logistic Regression, Random Forest (RF), K Nearest Neighbor (KNN), and Decision Tree have been applied for analyzing their performance which is based on the liver patient data. The dataset contains 583 liver patient's data whereas 75.64% male patients and 24.36% are female patients.

Keywords: Liver Disease, Machine Learning, Classification Algorithms, Logistic Regression, Random Forest, Decision Tree, K-Nearest Neighbor (KNN).

1. Introduction

The liver organ is bulky and brawny that placed on the right of the abdomen, and color is reddish-brown that feels rubbery to the touch. Normally can't feel the liver, because it's sheltered by the cage of ribs. The liver consists of 2 large sectors, named right and left lobes. The gallbladder situated under the liver, besides with element of the pancreas and intestines. The liver organ and its supporting organs provide assistance mutually to digest and process food. The liver gland is an incredibly significant organ that plays a very vital responsibility in the body and weight

approximately 3 pounds. It executes many tasks that help out to digest food, convert it into energy, and store the energy. It also assists to refine noxious substances out of the bloodstream. Diseases that may influence the liver include hepatitis, fatty liver, bleeding, fatigue, jaundice. When fat build-up around the liver known as fatty liver infection. Generally 2 major categories of fatty liver gland infection: an alcoholic fatty liver ailment which is originated by a huge quantity of alcohol consumption, smoking and poor lifestyle, and other one called nonalcoholic fatty liver ailment which is caused by other issues.

Difficulty with liver infection cannot find out until it is often too late as the liver continues to function even when partially damaged but sometimes this ignorance may be lethal. Early detection can potentially be life-saving but it is a tremendously difficult task; it needed highly proficient expertise, experience, and time. In spite of this, manual detection of liver disease has a high possibility of human error and error may lead you to compromise with human life. The aim of this paper is to apply classification algorithms to recognize liver patients from healthy people with the help of automated techniques, which is very useful for physicians in making an accurate judgment on patients and also gives a high level of precision rate with very less processing time. In medicinal diagnosis and predicting infections, classification methods play an extremely significant job. This is not mandatory a particular classifier is the preminent classification tool, the excellent performing algorithm will merely depend on features of the set of data is to be analyzed. In this work, four classification algorithms that are Logistic Regression, Random Forest (RF), K Nearest Neighbor (KNN), and Decision tree have been applied for analyzing their performance which is based on the liver patient data. In this test, we assemble a dataset from the UCI Machine Learning Repository. In addition, the original dataset was collected from the northeast of Andhra Pradesh, India. This dataset contains 583 liver patient's data whereas 75.64% male patients and 24.36% are female patients. This dataset has contained 11 exacting parameters whereas we choose 10 parameters for our further analysis and 1 parameter as an objective class.

2. Related Work

Sumedh Sontakke et al. [1] established a system which is dependent on enhanced diagnosis techniques of liver infection by exploring two methods of recognition- patient parameters and genome expression, the paper also discusses the computational algorithms that can be applied in the aforementioned methodology and lists demerits. Joel Jacob

et al. [2] recommend a system that is dependent on the classification algorithms to recognize liver patients from healthy persons. This project also goal to evaluate the classification algorithms based on their performance factors. To serve the medicinal community for the analysis of liver disease among patients, a graphical user interface will be developed using Python. Atrayee Dutta et al. [3] proposed a work that is proficient to identify the cancer cell of the liver in an MRI image. Here ostu's technique is utilized for enhancing and watershed scheme, for segmenting the cancer cell from the picture. Francisco Perdigón Romero et al. [4] established a system that is proficient to classify cysts and metastases from CT scan pictures via deep neural networks from mechanized segmentation. The planned model allows for the utmost accuracy and AUC compared to previous state-of-the-art approaches. The framework provides a screening tool for the initial stage finding of malignant tissue which can be combined into CAD software. P.Yugander et al. [5] this work shows an improved DRLSE model to segment hepatic lumps from the strident CT scan pictures. This approach is capable of handling noisy medical pictures, especially CT scan pictures. The main intention of this algorithm is to segment noisy medical images. Also proposed model, an evolving contour is results illustrate the efficiency of the recommended system, and the results achieved by MATLAB. Amita das et al. [6] proposed methodology is valuable for the involuntary segmentation of liver cancer on CT scan representations. This automatic extraction achieved using spatial fuzzy c-mean clustering. The main focus of this algorithm is to segment lumps and early diagnosis of cancer diseases that are helpful for clinicians. Hong Zhou et al. [7] proposed methodology which is based upon bendable metamaterial biosensor fabricated through sintering Ag NPs on ubiquitous. A scheme based upon the relationship b/w the resonance and permittivity of the bimolecular. This effort provides an approach of flexible, inexpensive, and unlabeled biological recognition,

potentially benefiting the growth of bioassay. Here the difficulty is identified with accuracy and behavioral approach that degrade the system towards lower accuracy

3. Proposed Work and Implementation

A. Dataset

In this test, we gather datasets from the UCI Machine Learning Repository [11]. In addition, the unique set of data was assembled from the northeast of Andhra Pradesh, India. This dataset contains 583 liver patient's data whereas 75.64% male patients and 24.36% are female patients. This dataset has contained 11 exacting parameters whereas we choose 10 parameters for our further analysis and 1 parameter as an objective class. The complete explanation of the dataset is specified via the chart. The chart (Table 1) depicts detail regarding the attribute, attribute type, and attribute description.

Table 1: Dataset Description

#	Attribute Name	Attribute Type	Attribute Description
1	Age	Numeric	Patient's age
2	Sex	Categorical	Patient's gender
3	Total Bilirubin	Numerical	Total bilirubin quantity in patient
4	Direct Bilirubin	Numerical	Direct bilirubin quantity in patient
5	Alkphos Alkaline Phosphotase	Numerical	ALP enzyme amount in patient
6	Sgpt Alamine Aminotransferase	Numerical	SGPT amount in patient
7	Sgot Aspartate Aminotransferase	Numerical	SGOT amount in patient
8	Total Proteins	Numerical	Protein content in patient
9	Albumin	Numerical	Albumin amount in patient
10	Albumin and Globulin Ratio	Numerical	Ratio of albumin and globulin
11	Class	Nominal	Liver disease status in patient

B. Data Pre-processing and feature selection

In this effort, we examine 583 liver patient's data whereas 416 samples are fatty liver patients and 167 samples are non-fatty liver patients. The part of the whole liver patients is shown in Figure.1. Moreover, the liver patient's dataset, shown in Figure 2, where 441 are male samples and 112 are female samples were taken for examination. The pre-processing method was applied to normalize the missing values. The missing values along with their instances were replaced by a null value. It was followed by feature selection to recognize significant attributes for categorization. Feature selection was performed using filter and wrapper techniques. Primarily, the attributes hold above 70% correlation was eradicated from the dataset by correlation inspection. The plot shown in Figure 3 illustrates the age group of infected populace.

Figure 1: Gender count in dataset

Figure 2: Gender ratio in diseased patients

Figure 3: Age group of infected populace

The heat map is shown in Figure 4 which appears to contain some correlated parameters. Some of these columns contain low correlation and were omitted.

Figure 4: Heat map to check correlated columns for the liver dataset

C. Classification Algorithms

(i) Logistic Regression (LR)

LR is an easy classification approach of supervised ML algorithm. It makes a prophecy that holds the binary result from the previous data. Generally, this algorithm provides the outcome in a very small instant. LR can deal with any number of numerical with absolute factors. In addition, it introduces a discrete parallel item somewhere in the variation of 0 and 1. Strategic Regression processes the correlation between the element factors by surveying probabilities (p) utilizing an underlying logistic function.

(ii) Random Forest (RF)

RF or Random Decision Forests are an assembly learning procedure for categorization, regression and special assignments that works by developing a huge no. of decision trees at instruction instant and yielding the set that is the approach of the classes or means forecast (regression) of the individual trees.

(iii) K-Nearest Neighbors (KNN)

KNN is a category of supervised ML algorithm which employed for grouping as well as regression predictive predicaments. KNN also store up each samples in the training sets and categories fresh samples depend on a majority of its neighbors. In this approach K is an arbitrary positive integer no. that represents the no. of neighbors around the unfamiliar sample. The selection of this parameter is essential and can influence the classification effort. The expanse between the new sample and the classified samples can be achieved via diverse approaches. The most frequent distance function is the Euclidean distance function. The Euclidean distance between points $(x_1, \dots, x_k) \wedge (u_1, \dots, u_k)$ is specified as equation (1).

$$E = \sqrt{\sum_{i=1}^k (x_i - u_i)^2} \quad (1)$$

(iv) Decision Tree (DT)

DT algorithm groups the data in the form of a tree configuration by breaking down the difficult decision making procedure into subsets of a simple decision. The approach has a root, internal and leaf node, in all of these the root node is the excellent predictor method in a tree. A DT algorithm can be helpful for taking care of regression and categorization issues as well. The general thought process of utilizing DT is to build a training model that used to predict class or estimation of objective factors by taking in choice standards derived from earlier data (training data). The approach split the population into 2 or more sets. The two phases are done recursively through the remaining features.

$$\text{Entropy}(S) = \sum_{i=1}^c -p_i \log_2 p_i \quad (2)$$

Let it be more defined via flow graph of proposed idea (Figure 5). The similar procedure will be followed as described above.

Figure 5: System Architecture

4. Result Analysis

To evaluate the performance of the classifiers and as a systematic comparison approach, six typical parameters of accuracy, precision, recall, specificity, F-measure and FPR are implemented. These parameters are broadly being used and an integration of their values can act as a proper indicator of the performance of a classification algorithm. We gather dataset from the UCI Machine Learning Repository [11]. In addition, the unique set of data was assembled from the northeast of Andhra Pradesh, India. This dataset contains 583 liver patient's data whereas 75.64% male patients and 24.36% are female patients. This dataset has contained 11 exacting parameters whereas we choose 10 parameters for our further analysis and 1 parameter as an objective class.

(i) Accuracy: The precision of the classifier is percentage of data points that are appropriately classified through the classifier.

$$\text{Accuracy} = \frac{\text{No. of TP} + \text{No. of TN}}{\text{No. of TP} + \text{TN} + \text{FP} + \text{FN}} \quad (3)$$

False Negative (FN) = no. of records foreseen to be negative that are really positive.

True Positive (TP) = no. of records foreseen to be positive that are really positive.

False Positive (FP) = no. of records foreseen to be positive that are truly negative.

True Negative (TN) = no. of records foreseen to be negative that are really negative

(ii) Sensitivity or Recall: Sensitivity is the ratio of TP that are appropriately classified. It is also recognized as the true positive rate (TPR).

$$\text{Sensitivity} = \frac{\text{No. of TP}}{\text{No. of TP} + \text{No. of FN}} \quad (4)$$

(iii) Precision: It is the percentage of TP outcome vs. every positive outcome

$$\text{Precision} = \frac{\text{No. of TP}}{\text{No. of TP} + \text{FP}} \quad (5)$$

(iv) Specificity: it is the proportion of negative tuples that are appropriately recognized.

$$\text{Specificity} = \frac{\text{No. of TN}}{\text{No. of TN} + \text{FP}} \quad (6)$$

(v) F-measure: It the precision of the model by a blend of accurateness and recall.

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

(vi) FPR: A false positive ratio also called false alarm ratio is the possibility of wrongly declining the null hypothesis for a meticulous test. The FPR is estimated as the proportion between the no. of negative occurrences incorrectly classified as positive (false positives) and the entire number of actual negative actions.

$$\text{FPR} = 1 - \text{Specificity} \quad (8)$$

The Confusion Matrix for Applied classification model is presented in Table 2
The consequence of every classification algorithms is sum up in the tabular form in Table 3

Table 2: Confusion Matrix for Applied classification model

Classification Models	Terms			
	TP	TN	FP	FN
Random Forest	285	94	154	50
Logistic Regression	298	111	112	62
Decision Tree	298	72	116	97
K-Nearest Neighbor	395	25	155	8

Graph No. 1 Attributes comparison of the applied classification models.

Figure 6: comparison of the applied Classification models.

Table 3: Consequences of classification algorithms

Classification models	Accuracy (%)	Sensitivity /recall (%)	Specificity (%)	Precision (%)	F-Measure (%)	RMSE (%)
Random forest	65.00	85.07	37.90	64.92	73.64	59.039
Logistic Regression	70.15	82.77	49.77	72.68	77.39	54.51
Decision Tree	63.46	75.44	38.2	71.98	73.66	60.474
K-Nearest Neighbor	72.04	98.01	13.8	71.81	82.88	52.915

5. Conclusion

In this framework, we proposed a technique for identifying liver disease using a machine learning algorithm which is capable enough to improve predictive power. The four most

operable machine learning methods that were used include Logistic Regression, Random Forest (RF), KNN, and Decision tree. The system was implemented with all the models and their performance was estimated. K Nearest Neighbor (KNN) was the model that

generated a great accuracy rate with an exactness of 72.04%. This is not mandatory a particular classifier is a great classification tool; the excellent performing algorithm will merely depend on features of the set of data to be examined. The benefit of the established scheme is that it has high accuracy and also able initial phase recognition of liver disease automatically to overcome human fault and reduce false alarm rate that gives the progression which is additionally accurate, effortless, and unproblematic. In the future, we expand our research on diverse data set to depict a more general conclusion and eradicate misclassification, as false recognition in medicine could not be accepted.

References

1. Sumedh Sontakke¹, Jay Lohokare, Reshul Dani, Diagnosis of Liver Diseases using Machine Learning, Proceedings of the International Conference on Emerging Trends & Innovation in ICT, Pune, India, 2017.
2. Joel Jacob¹, Joseph Chakkalakal Mathew, Johns Mathew, Elizabeth Issac, Diagnosis of Liver Disease Using Machine Learning Techniques, International Research Journal of Engineering and Technology, Vol. 05 Issue 04, Apr-2018. 4011-4014.
3. Atrayee Dutta and Aditya Dubey, Detection of Liver Cancer using Image Processing Techniques, Proceedings of the International Conference on Communication and Signal Processing (ICCSP 2019), 4-6 April 2019, Melmaruvathur, Tamil Nadu, India, 315-318.
4. Francisco Perdigón Romero, Andre Diler, Gabriel Bisson-Gregoire, End-To-End Discriminative Deep Network For Liver Lesion Classifications, 2019 Proceedings of the IEEE 16th International Symposium on Biomedical Imaging, April 8-11, 2019, Venice, Italy, 1243-1246.
5. P.Yugander and G.Raghotham Reddy, Liver Tumor Segmentation in Noisy CT Images using Distance Regularized Level Set Evolution based on Fuzzy C- Means Clustering, Proceedings of the IEEE 2nd International Conference On Recent Trends in Electronics Information & Communication Technology, Bangalore, India, May 19-20, 2017.
6. Amita das and priti das, Detection of Liver Cancer Using Modified Fuzzy Clustering and Decision Tree Classifier in CT Images, 2019 Pattern Recognition and Image Analysis, volume 29, 2019, 201–211
7. Hong Zhou¹, Cheng Yang³, Donglin Hu¹, Flexible Nano-Ag@paper Biosensor and its Application in Detection of Liver Cancer, 2018 Proceedings of the IEEE 13th Annual International Conference on Nano/Micro Engineered and Molecular Systems, Singapore, 22-26 April 2018.
8. M. Banu Priya, P. Laura Juliet, P.R. Tamilselvi, Performance Analysis of Liver disease prediction using Machine Learning Algorithms, International Research Journal of Engineering and Technology, Volume 05, Issue 01, January 2018. 206-211.
9. N. Li, Y. Jiang, G. Gong, G. Han and J. Ma, Non-Invasive Assessment Model of Liver Disease Severity by Serum Markers Using Cloud Computing and Internet of Things, IEEE Access, Vol. 6, 2018, 33969-33976.
10. S.Vijayarani, S.Dhayanand, Liver Disease Prediction using SVM and Naïve Bayes Algorithms, International Journal of Science, Engineering and Technology Research, Volume 4, Issue 4, April 2015, 816-820.
11. UCI Machine Learning Repository. ILPD(Indian Liver Patient Dataset) <https://archive.ics.uci.edu/ml/datasets/ILPD> (Indian+Liver+Patient+Dataset).