Examining The Risk Factors of Liver Disease: A Machine Learning Approach

Conference Paper · June 2022

DOI: 10.1109/ICCES54183.2022.9835732

Md. Sagar Hossen
Institut Teknologi Sepuluh Nopember
21 PUBLICATIONS **290** CITATIONS
SEE PROFILE

Imdadul Haque
Daffodil International University
7 PUBLICATIONS **44** CITATIONS
SEE PROFILE

Puja Rani Sarkar
Daffodil International University
6 PUBLICATIONS **59** CITATIONS
SEE PROFILE

Md. Ashiqul Islam
Daffodil International University
29 PUBLICATIONS **390** CITATIONS
SEE PROFILE

Proceedings of the Seventh International Conference on Communication and Electronics Systems (ICCES 2022)
IEEE Xplore Part Number: CFP22AWO-ART; ISBN: 978-1-6654-9634-6

Examining The Risk Factors of Liver Disease: A Machine Learning Approach.

Md. Sagar Hossen [1]
Dept. of Informatics Engineering.
Institut Teknologi Sepuluh Nopember.
Surabaya, Indonesia.
sagar15-1504@diu.edu.bd

Md. Ashiqul Islam [2]
Dept. of Computer Science and Engineering.
Daffodil International University.
Dhaka, Bangladesh.
ashiqul15-951@diu.edu.bd

Imdadul Haque [3]
Dept. of Computer Science and Engineering.
Daffodil International University.
Dhaka, Bangladesh.
imdadul15-1440@diu.edu.bd

Wasik Ahmmed Fahim [4]
Dept. of Computer Science and Engineering.
Daffodil International University.
Dhaka, Bangladesh.
wasik15-1156@diu.edu.bd

Puza Rani Sarkar [5]
Dept. of Computer Science and Engineering.
Daffodil International University.
Dhaka, Bangladesh.
puza15-1262@diu.edu.bd

Tania Khatun [6]
Dept. of Computer Science and Engineering.
Daffodil International University.
Dhaka, Bangladesh.
tania.cse@diu.edu.bd

2022 7th International Conference on Communication and Electronics Systems (ICCES) | 978-1-6654-9634-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCES54183.2022.9835732

***Abstract*— Nowadays, Liver Disease (LD) is a very common clinical problem for human health and is related to morbidity and mortality. Nevertheless, an earlier prognosis of LD patients gets a scope to avoid, prior diagnosis and subsequent treatment. This research work attempts to implement a high qualified performer machine learning design to predict LD. the most wanted and unwanted risk factor of LD which could help physicians in classifying risky patients and create an analysis to restrict and control LD. The proposed research study has included all patients. who were identified as having liver diseases. Totally, 6 (six) machine learning algorithms such as Decision Tree(DT), Logistic Regression(LR), Multilayer Perceptron(MLP), Artificial Neural Network(ANN), Random Forest(RF), K Nearest Neighbor classifier(KNN) are selected to predict LD. The location underneath had been utilized to evaluate the accuracy among the six applied models. An overall total of 583 instances had been included in this scholarly research; of the 416 patients are affected by liver illness. The location which defines the receiver operating characteristic (AU ROC) of Logistic Regression. Decision Tree. Multilayer Perceptron. Random Forest. Artificial Neural Network. and K-Nearest Neighbor classifier with 10-fold-cross validation was performed. Furthermore, the reliability of LR, DT, MLP, RF, ANN and KNN with accuracy 72.89%, 81.32%, 60.24%, 86.14%, 75.61%. and 65.52%. The utilization of woodland which is certainly arbitrary within the medical setting may help doctors to detect and classify liver patients for major avoidance, surveillance, quick treatment, and management. LR, DT, MLP, RF, ANN, and KNN formulas are acclimatized to forecast and after analyzing the data set, an increased price of accuracy is achieved.**

Keywords— Early Stage, Liver Disease, Artificial Neural Network, Logistic Regression, Random Forest.

I. INTRODUCTION

Liver Disease (LD) is the foremost fatal illness worldwide and has become a crucial universal hygiene concern. The spectrum of LD ranges from simple Inflammation, Fibrosis Cirrhosis, End-stage disease (ESLD), and cancer of the liver. In a few years, the outbreak and fatality of affliction are rising, and therefore the final cause of affliction becomes much more difficult [1]. Liver function has disintegrated to the purpose that the damage can not be reversed aside from the liver transplant [2]. Therefore, early prediction of disease is of utmost importance to tackle the disease.

Effective treatment of the disease depends on the first recognition of varied risk factors. Studies show LD is related to multiple risk factors. These factors contain multiple variables like age, TB, DB, TP, SGOT, and AL alongside disease caused by different types of viruses, alcohol use, and obesity which are caused by unhealthy lifestyle and bad food habits. These cause massive data generation and decoration to guide the partitions for an informed decision about the LD patient.

There are 4 common early stages of liver disease. 1. Inflammation (early stage) 2. Fibrosis 3. Cirrhosis 4. End-stage [3]. If people are aware of the first stage of liver disease by its symptoms then it can be protected easily. The liver is a disease where 75% of the liver tissue has to be affected [4]. But at an early stage, this disease can't be identified easily. On the other hand, the treatment is lengthy and risky. And the risk factor for this disease like avoiding alcohol, being safe from viruses, and maintaining a healthy diet [5]. So if the condition of a patient can be predicted then they can recur easily. For this prediction machine learning is the best way in this modern technology to classify diseases [6], [7]. There has been used a machine learning approach to patient data to get the best accuracy on early-stage and risk factors. The blessing of machine learning can be seen in the medical sector too. The machine learning approaches are mainly used for predicting stages and risk factors. Herewith risk factors, we will identify the main reason or symptoms that can cause liver disease.

Nowadays, IoT and IoMT based health monitoring system is developed to detect the disease easily in the early stage [8] [9]. Computer Vision can help the medical doctor and physician to identify the disease using image processing and cure the disease of the root [10]. In this paper, ML algorithms are used on patient data to urge the simplest accuracy on the early-stage and risk factors. The ANN, Random forest, Perceptron, Logistic regression, and Decision Tree algorithms are performing for better accuracy. The data set was culled and run the test successfully and it provides the great results.

II. LITERATURE REVIEW

Machine learning is now a blessing to modern technology. It reduces time and effort, especially in medical science. Once identifying diseases was a very lengthy or time-consuming process. Some of them were costly too. But with the help of machine learning techniques is very easy to make the identification or predict any disease. Researchers perform feature selection methods, machine learning techniques, and subset selection sequentially to analyze liver disease [11].

The author [12] applied a feature selection method to classify the predictive model and perform the J48 algorithm with an accuracy rate of 95.04%. The source of their data is the UCI repository. The authors [13], [14] used the feature selection method to predict the liver disease risk level. There are various classifying algorithms which are Logistic Regression, Random Forest, SMO, Naive Bayes, k-nearest neighbor (Ink) and J48 are implemented on the Liver Patient dataset to get the acceptable accuracy. Subsequently, the study showed that LR gives the highest accuracy of 72.50% and RF gives 71.53%. They [2] used Back Propagation or ANN and the KNN algorithm is applied with different feature combinations. Overall, the study showed that back Propagation was 98.002 (Accuracy) 99.4 (Precision) 92.22 (Sensitivity) 99.82 (Specificity). The liver tissue needs to be affected by 75% to decrease the functionality of the liver. Classification algorithms are Naïve Bayes classifier, C4.5, Backpropagation Neural Network algorithm, and Support Vector Machines, and these are evaluated based on mainly four criteria as Accuracy, Precision, Sensitivity, and Specificity. In [15] author used KNN and LR to represent the flow of data processing with feature selection. The accuracy of both KNN and LR is 73.97%.

Authors [16] described the parameter of different classification algorithms for predicting liver disease. He classified RF, KNN, DT, Adaboost, and Gradient Boost (GB) Classifier to early predict liver disease and achieve the highest accuracy from RF at 77.2%. In [17] author used 6 different algorithms (KNN, LR, and RF, DT) for predicting liver disease. The corresponding accuracy of these algorithms is 62%, 75%, 74%, and 69%. Researchers [18] classify chronic liver disease with Machine Learning Algorithms. The highest accuracy achieved from the SVM algorithm is 99.76 % and Boasted C5.0 model has provided the maximum precision of 97.52% and the K-means model has the highest sensitivity at 97.36%. Researchers [19] can apply KNN and SVM to predict and analyze liver disease via a machine learning model. They [20] are applying a 2-years survival analysis using child Pugh and MELD score to analyze the liver and cirrhotic disease patients. Researchers [21] can apply inclusion criteria, IL-6, MPV, and an abdominal ultrasound too early detection of fatty liver disease. They collected the data from Prof. Dr. R. D. Kandou Manado's general hospital. Researcher [22] diagnoses the challenges of pregnancy in liver disease. They have evaluated liver enzymes, hemolysis, preeclampsia, and low platelet count syndrome to diagnose the disease.

They [23] proposed a method of identifying liver disease in high-risk patients, diagnosis, prevention, and management. They performed NB, ANN, LR, RF, and RF showing the highest accuracy. They [24] emphasized using subset selection for improving performance with the highest accuracy. Author [28] mentioned an efficient way of resembling different feature subsets generated by the random subset method. To predict liver disease author culled a dataset from the UCI repository [15]. The author used Random forest and other algorithms and RF provides the second-highest accuracy of 70.38% for selecting feature subset. Authors [11] show the occurrence of liver disease and mentioned 100 types of infection. They [12] discussed the limitations of the ultrasound technique. They [13] reported shows some algorithms with a confusion matrix. Detecting liver disease from early-stage and with proper diagnosis patients can fully recover. So, the author [14] believed it can bring down the cost by about 30%. Authors [15] implemented a feature model for predicting liver disease in 3 phases with dataset attributes. Symptoms of liver disease and predicting them through different algorithms are mentioned. Researchers[1] talk about big data and clinical information for disease datasets. Author [25] reported that Liver disease can be associated with cardiovascular disease. Also, he mentioned some attributes to use. He [26] mentioned the reason for liver disease and also shows some algorithms with attributes to evaluate accuracy. Author [27] proposed a work to predict liver disease through classification algorithms. He [28] represents the major three liver disease predictions with classification algorithms. The decision tree was used to evaluate accuracy. Different types of attributes were mentioned by the authors [29]. In [2], the author uses 5 different classification algorithms where KNN and DT were mentioned. The researcher [30] represents the use of

Classification and Regression Tree (CART). He also said that many medical applications used it because it has greater classification compatibility and a rear method of including recent AI in disease prediction [31]. Researchers [32] identify the machine learning technique and treatment process of covid-19 pre-existing affected liver injury. Researchers [33] culled data from the Isfahan Cancer Registry to calculate 100,000 people's data to analyze the period prevalence (PP) and incidence rates (Irs) for liver cancer. They [34] analyze the gene expressions of IR and NAFLD to occur early prediction of liver disease. Author [4] represents the importance of classification algorithms choosing techniques. He said Normalizing classification data can improve the overall diagnostic accuracy.

Authors [35] have proposed a machine learning approach to an early prediction model to predict CAD (Coronary Artery Disease) as a comparative study because multiple algorithms have been used to predict. Naive Bayes and SVM machine learning algorithms provide different accuracy with the medical datasets. Researchers [36] have implemented a feature extraction analysis of ECG signal interpretation based on Fusion and its systematic approaches. There have been used different algorithms with different accuracy but Random Forest was a performance more effective to make the ECG classification with 350 samples of the dataset for training and testing.

Various approaches have been published related to liver disease but failed to point out individual symptoms that are more dangerous for the human body. For the symptoms impact of liver disease, our proposed models find out a perfect solution by risk factor and this is the main reason for the proposed model.

TABLE.1. Comparison between Existing Model and Proposed Model

Existing Model	Proposed Model
Maximum existing models didn't compare with different algorithms [29].	There are different types of algorithms that have been applied to find a better solutions
existing models didn't show the precision, recall, and f1 score result which can able to find a better performer approach	There have been shown the precision, recall, and f1 score in the proposed model.
There are no specific symptoms discussed and symptoms mostly destroy our health stability.	It clearly explained the most dangerous symptoms for the human body.
There is no explanation for the risk factors for the liver disease symptoms [17].	The risk factor has been shown to identify the most dangerous symptoms for a healthy life.

III. PROPOSED MODEL

The Liver disease classification using machine learning approaches is implemented and we have used there are six isolated algorithms as ANN, LR, RF, DT, MLP, and KNN. Our main theme is to make the identification of risk factors of LD to detect the disease in the early stage.

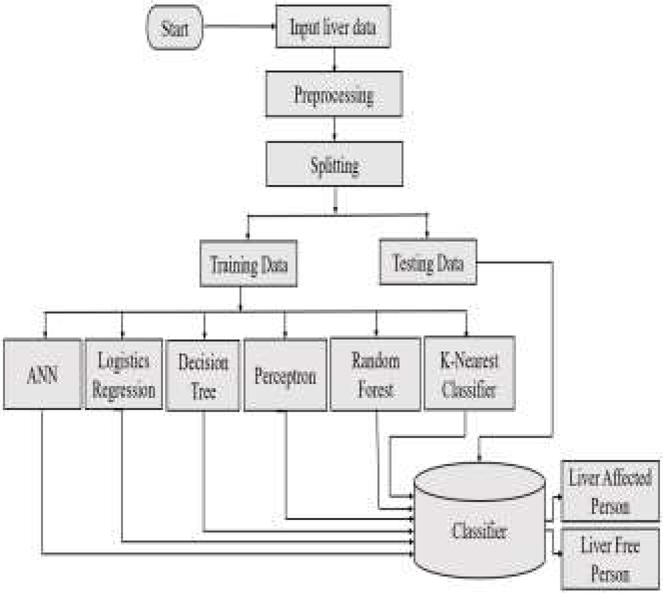


Fig.1. Proposed Model

From Fig.1, we can define our whole model of LD classification using various machine learning models. We have used the real-life dataset as stored in a CSV file and first of all, we took the data from our liver dataset as input and then needed to preprocess the dataset, and preprocesses mean the need to remove null values if there are. The splitting part mainly is completed by using the sklearn model and we have taken 20% of liver data is used for testing and 80% of data is used to train the model. After successfully training the 80% data using the different six types of machine learning and deep learning model made a classification of the trained model with the respect to 20% test data of the liver. For the different types of algorithms, we have gotten different accuracy as 60.24% is for MLP, 65.52% is for the KNN, 72.89% is for LR, 75.61% is for ANN, 81.32% is for DT and 86.14% is for the RF. These algorithms can ensure that the Random Forest machine learning model performs better than another algorithm [27]. That means the liver disease patients can use a random forest algorithm for predicting liver disease. And our model is implemented for liver disease so that they can easily check their disease concerning covetable symptoms.

IV. DATASET DESCRIPTION

The dataset is collected from UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset) is a multivariate dataset with 583 instances [2]. 10 attributes are used in the dataset to collect the patient data. The data attribute is described in Table-2.

TABLE.2. Dataset attribute Description

Attribute	Description	Values/measurement
Age	It refers to the age of every patient in a year.	numeric number
Gender	It refers to the gender of patients in the binary category	• 0: means Male • 1: means Female
TB	It refers to the Total Bilirubin of each patient.	• 0.4 mg/l to 74 mg/l in numeric value.
DB	It refers to Direct Bilirubin in patients.	• 0.1 mg/l to 19.7 mg/l in numeric value.
Alkphos	It refers to Alkaline Phosphatase.	• The normal range is 44 to 147 international units per liter (IU/L). • Data set: 63 to 2110.
Sgpt	Alanine Aminotransferase.	• The typical scope of qualities for ALT (SGPT) is around 7 to 56 units for each litter of serum. • Data set value : 10 to 2000 u/l.
SGOT	It refers to Aspartate Aminotransferase and the typical scope of an SGOT.	• SGOT range is 8 and 45/ltr. • In the dataset, there rage 10 to 4929 IU/L.
TP	It refers to Total Proteins.	• Total Proteins in the typical reach are 6.0 to 8.3 g/dL. • We have 2.7 to 9.6 g/dL among all patients.
ALG	It refers to Albumin in the patient's body.	• Normal range is 3.4 to 5.4 g/dL • Data set: 0.9 to 5.5 g/dL.
A/G	Ratio Albumin and Globulin.	• Normally, there is a little more albumin than globulins, giving a normal A/G ratio of slightly over 1. • We got 0.3 to 2.8 g/dL
Class	The selector field is used to split the data into two sets (labeled by the experts)	Disease not present

A. Data Preprocessing

After collecting data the main work is data processing before the training model. Systematically, we have to preprocess the data [37]. First of all, we describe the data and then check the null value of each attribute. After finding that we got some null value then we fill up the blank using the null value handling method mean, median [29]. We use the mean value to fill up the blank of each attribute. We check the boundary error of our data. There was no boundary error. we define the splitting steps separately and the included steps are like data occurring, importing the data, checking the missing value, and handling the missing value. Then encoding the categorical data. Then converted all of the data into an integer so that it was ready to split the dataset into train and test datasets. Then we got the cleaned encode data. After that, we split the dataset and feature scaling which is a technique to normalize the free factors of a dataset inside a particular reach. As such, highlight scaling limits the scope of factors with the goal that you can analyze them on normal grounds. It is the ending part of preprocessing.

V. Model Description

ANN: After collecting data the main work is data processing before the training model. Systematically, we have to preprocess the data. First of all, we describe the data and then check the null value of each attribute. After finding that we got some null value then we fill up the blank using the null value handling method mean, median [29]. We use the mean value to fill up the blank of each attribute. We check the boundary error of our data. There was no boundary error [38]. Then encoding the categorical data. Then we got the cleaned encode data. After that, we split the dataset and feature scaling which is a technique to normalize the free factors of a dataset inside a particular reach. As such, highlight scaling limits the scope of factors with the goal that you can analyze them on normal grounds. It is the ending part of preprocessing. ANN algorithm is modeled like a human brain, as it can learn data, similarities, and responses from past data. Also, it can respond to predictions. ANN model has 3 layers and that's are input, hidden and output layer. The required equation for the ANN algorithm is given below, where a, is the output and "p" defines the input. Parameters "w" and "p" specifically defines the weight and bias of that equation:

$$a = f(net) = f(n) = f(w^T.p + b) = f(\sum_{i=1}^{R} w_R^T.p_R + b)$$
(1)

Multilayer Perceptron: A perceptron is a neuron's computational prototype that is categorized in form of a neural network. A perceptron algorithm has one or higher than one input but only one output. This algorithm used a linear classifier for making it easy for a binary classifier. The purpose of this algorithm is to take correct labels of data for prediction or training models. From the bestowed equation "y" defines the output and w defines the weight, "x" defines the input and θ defines the angle of "w" and "x".

$$y = 1 if \sum_{i=1}^{n} w_i * x_i - \theta \geq 0 \qquad (2.1)$$

$$y = 0 if \sum_{i=1}^{n} w_i * x_i - \theta < 0 \qquad (2.2)$$

K-NN: K-NN means K-Nearest Neighbor, which defines a supervised machine learning approach. K-NN finds the similarity between new data with available data and then puts the new case in the most available categories. On the other hand, K-NN is used not only for regression but also as classification algorithms. But it is mainly used as a classification algorithm. During the training phase, K-NN stores datasets, and when new data are given it classifies the data and puts it into the most similar category. For calculating the K-NN algorithm first we have to find the nearest distance through distance functions: Euclidean, Manhattan, Minkowski. For an example equation for Euclidean distance function:

$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} \quad (3)$$

After finding the distance function the algorithm will take several neighbors depending on the K value. Then the algorithm will put the data in the most matched category.

Decision Tree: Decision tree algorithms are applied to solving the issue of regression, and classification type of problems. The main objective of this algorithm is mainly used to place the class or value of the target variable by learning the simple decision rules which are inferred from training data [39]. A Decision tree fundamentally poses an inquiry of yes or no dependence on itself further splitting the tree into sub-tree [25].

The decision tree defines two hubs root and sub-hubs. While implementing a decision tree we use ASM (Attribute Selection measure) because of the attribute selecting problem between the root hub and sub-hubs. For ASM there are two popular techniques.

1. **Gini index:** it's a measurement of impurity or purity used in a CART (Classification and Regression Tree) decision tree model. The attribute of Gini index rules should be preferred compared to low to high. It creates the binary split s and CART algorithm using the Gini index via creating binary splits. It can be calculated as:

$$GiniIndex = 1 - £jPj2 \quad (4)$$

2. **Information gain:** It's an estimation of the changes in entropy after a division of the dataset dependent on the attribute. Entropy is a metric that actions the pollution of an attribute. Essentially, it determines irregularity in data. Entropy can be calculated as:

$$Entropy(s) = -p(yes)\,log2p(yes) - p(no)\,log2p(no) \quad (5)$$

Where is the total number of samples? P (yes) is the probability of the yes and p (no) is the probability of the no. And, the information gained can be calculated as:

$$Informationgain = Entropy(S) \quad (6)$$

Logical Regression: Logical Regression is mainly a classification model. It is used for president binary outcomes. It fits the data to build a logic function for predicting the probability of occurrence of an event. Logical regression can be calculated as:

$$g(E(y)) = \alpha + \beta \times 1 + \gamma \times 2 \quad (7)$$

When g is a link function, E(y) is the expectation of the target variable, and α, β, γ is a predictor.

The equation for Logistic regression with dependent variable enclosed in a link function is:

$$g(y) = \beta o + \beta \quad (8)$$

In this function, there is a two thing probability of success (p) and the probability of failure (1-p). For any value of slope and different variables, the exponent of this equation will never be negative.

$$P = exp(\beta o + \beta) = e^{(\beta o + \beta)} \quad (9)$$

Random forest: Most popular machine learning algorithm is Random forest becomes the most popular machine learning for good and ease of use. RF algorithm is used for classification or regression and other tasks which is operated by constructing a multitude of decision trees at training time. It is outputting mode of the classes or means or average prediction of individual trees.

The random forest algorithm equation can be calculated as:

$$MSE = \frac{1}{N}\sum_{i=2}^{N}(fi - yi)^2 \quad (10)$$

When N is the number of the total data points, fi defines the returned value by the applied model, and yi defines the actual value for the total data point i. Random forest is mainly performed based on the classification of data [40]. So there is mainly have to use the Gini index and the Gini index is calculated as:

$$Gini = 1 - \sum_{i=1}^{C}(p_i)^2 \quad (11)$$

Now, Gini defines the class, the probability to recite the specific branch of the nodes and pi determines the relative frequency of the class.

After reading the data with the help of the python pandas library, it checked if there is any null values by the isnull() function and then used the dropna() method to remove the null values if there were any. Then applied map() function to define males and females by 0 and 1 because of needing the numeric value to split the whole dataset into train and test.

After splitting the dataset, it is ready to apply a random forest algorithm to predict liver disease.

Firstly we import RandomForestClassifier from 'sklearn. esemble' and then fit the model via splitting the dataset as x_train and y_train. Finally, we made the accuracy compared to the trained model and test dataset and then it performed with the height accuracy of 86.14%.

VI. Experimental Analysis

This paper contains six different machine learning and deep learning algorithms to predict the performance of liver disease data.

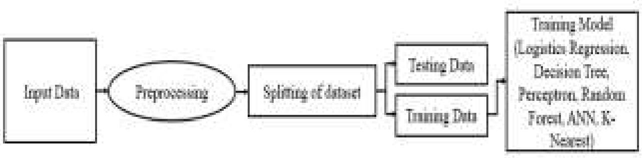

Input Data
Preprocessing
Splitting of dataset
Testing Data
Training Data
Training Model (Logistics Regression, Decision Tree, Perceptron, Random Forest, ANN, K-Nearest)

Fig.2. Training Model concerning Required Algorithms

First of all, collecting our liver disease data and setting the data as input data may have some null and unwanted data that's why we have to remove means clean all unwanted and null data as preprocessing of the liver data. After preprocessing we train the model and split the liver dataset as a training and testing dataset. Training data are used for training our model where we have used six different types of ML and DL algorithms [31] LR, DT, perceptron, RF, ANN, and k-nearest classifier algorithms. Now our model is ready to perform with 20% of the testing data.

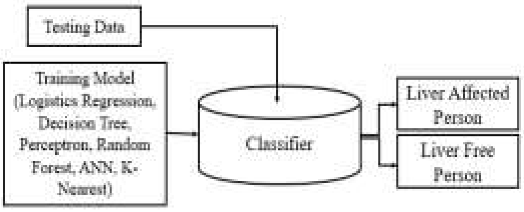

Testing Data
Training Model (Logistics Regression, Decision Tree, Perceptron, Random Forest, ANN, K-Nearest)
Classifier
Liver Affected Person
Liver Free Person

Fig.3. Classification

The model made the classification with testing data and provides the result as "liver disease affected person" or "liver disease-free person". From the ML and DL algorithm, we have gotten various confusion matrices to predict the performance of the algorithm individually which helps to measure the acceptance of the algorithm concerning the accuracy, precision, recall, and f1 score.

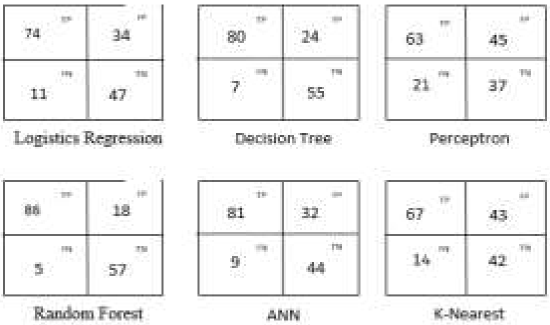


Fig.4. Confusion matrix

Now, we are going to use the below equations for finding accuracy, precision, recall, f1 score from figure-5 (confusion matrix),

TABLE.3. Accuracy, Precision, Recall, F1 Score

Algorithm Name	Accuracy	Precision	Recall	F1 Score
Logistics regression	72.89%	68.52%	87.06%	76.69%
Decision Tree	81.32%	76.92%	91.95%	83.77%
Perceptron	60.24%	58.33%	75%	65.62%
Random Forest	86.14%	82.69%	94.51%	88.07%
ANN	75.61%	71.68%	90%	79.80%
K-Nearest Classifier	65.52%	60.91%	82.71%	70.16%

From the above Table-3, there are different accuracy and f1 score for the different algorithms as logistics regression's accuracy and f1 score is 72.89% and 76.69%, for decision tree accuracy and f1 score is 81.32% and 83.77%, for perceptron accuracy and f1 score is 60.24% and 65.62%, for random forest accuracy and f1 score is 86.14% and 88.07%, for ANN accuracy and f1 score is 75.61% and 79.80%, fork-nearest classifier accuracy and f1 score is 65.52% and 70.16%. So, from the accuracy and f1 score, it is proved that random forest algorithms perform better than other machine learning and deep learning as ANN which helps doctors and patients easily check the danger of the liver disease of the human.

VII. Result Analysis

This is mainly comparison research where we have used six variant types of machine learning and deep learning algorithm with different accuracy and now, I need to find the AUC for making a better comparison within six different algorithms.

TABLE.4. Accuracy, AUC

Algorithm Name	Accuracy	AUC
Logistic Regression	72.89%	72.54%
Decision Tree	81.32%	80.79%
Perceptron	60.24%	60.06%
Random Forest	86.14%	85.25%
ANN	75.61%	73.95%
K-Nearest Classifier	65.52%	66.06%

Table- 4, defines the difference between accuracy and AUC value where 86.14% is the most acceptable accuracy and 85.25% is the best means for larger AUC which is for the Random Forest model. Random Forest performs better than the other ML and DL algorithms and RF is the most acceptable fit for the implemented dataset. It also finds a risk factor using a linear regression algorithm. The factor analysis result is shown in Table-5.

TABLE.5. Risk Factor Analysis

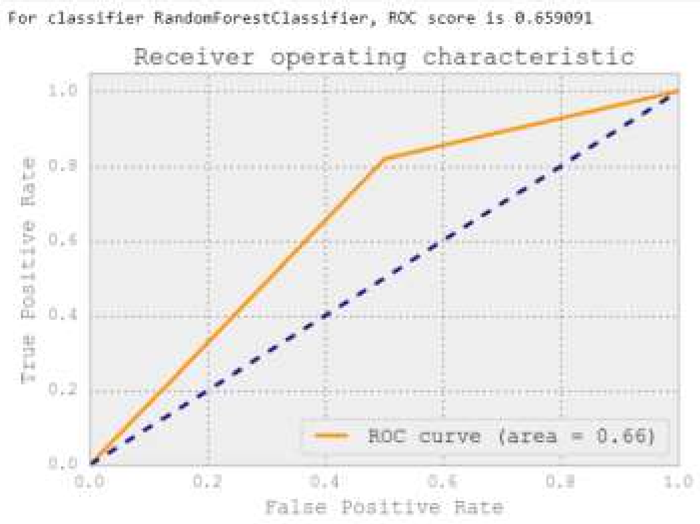
Algorithm Name	Factor
Linear Regression Model	-0.0032 * 65 (Age) + -0.0242 * 0.1(DB) + -0.0002 * 187 (Alkphos) + -0.0003 * 16 (Sgpt) + -0.0591 * 6.8 (TP) + 0.1069 * 3.3 (ALG) + 1.5916

In this study, we are applying the Linear Regression algorithm mainly to analyze the effective and non-effective risk factors of liver-affected patients [3]. All the required liver patient data was collected from the UCI repository. Basically, in using data with different attributes such as Age, Gender, TB, DB, Alkphos, Sgpt, shot, TP, ALG, A\G, and Class. Depending on this type of attribute and the range of the human body, machine learning models as linear regression can find out the risk factor of LD within a very short time.

The most effective reason to find the most dangerous symptoms which are harmful to a healthy life. The linear regression model has been used to find the risk factor to apply to liver disease attributes. In this approach, liver disease datasets are split into training and testing, and then apply linear regression algorithms to train the model. After the training model, we evaluate the risk factor based on every attribute of the liver disease dataset and we successfully did it in Table-04.

From the result analysis, we can see that Total protein was the most significant risk factor for liver disease. Because proteins are very useful in the human body. Age is another reason for liver disease as is shown in Table-5. Older are more vulnerable to liver disease. Besides ALG seems to behave the less effect the causing liver disease. So from the table, it can be defined the total protein is defined as significant, and ALG defines the way of a non-significant factor in liver disease.

Checking the missing value and handing the missing values and then converting those strings into an integer so that we can easily split the dataset into train and test. The most important thing to increasing the height accuracy from the random forest algorithm is to import RandomForestClassifier from 'sklearn. ensemble'.



```
For classifier RandomForestClassifier, ROC score is 0.659091
```

Fig.5. ROC Curve

ROC curve mainly defines the graphical features as the performance of the classification model which obeys the trade-off rules mean one axis increases and another axis must decrease. From the above ROC curve graph, False Positive Rate (FPR) and True Positive Rate (TPR) define the x-axis and y-axis. When the value of the x-axis (FPR) increases, then the y-axis (TPR) decreases, and also happens the increasing and decreasing in the opposite. The score of the ROC curve is 0.66 and the score between 0.50-1.0 is known as a better performance of the model Random Forest is also defined as a more accurate performance as a ROC curve with pretty accuracy.

ROC curve is mainly used for evaluating the metrics to show graphically for binary classifications. From the confusion matrix, we get the value of True positive, True Negative, False Positive, and False Negative.

True Positive Rate = True Positives / (True Positives + False Negatives) (12)

Sensitivity = True Positives / (True Positives + False Negatives) (13)

False Positive Rate = False Positives / (False Positives + True Negatives) (14)

Specificity = True Negatives / (True Negatives + False Positives) (15)
False Positive Rate = 1 - Specificity (16)

Concerning equations (12), (13), (14), (15), and (16) we can get the value of the x-axis and y-axis which make a graphical representation like the ROC curve.

VIII. COMPARATIVE ANALYSIS

As discussed before in the Literature review, there is enormous work has been done with liver disease using machine learning and deep learning based on real-life data.

TABLE.6. Comparisons

Studies	Methods	Accuracy	Datasets
Jag deep Singh et al. [13]	LR, RF	72.50%, 71.53%	UCI Repository
A.K.M Sazzadur Rahman et al. [17]	LR, RF, DT, KNN.	75%, 74%, 69%, 62%	UCI Repository
M. Phill et al. [41]	RF with PSO feature selection and Greedy Step Wise	70.32%, 80.22%	UCI Repository
Sumedh Sontakke et al. [42]	ANN Back Propagation	73.2%	UCI Repository

Some important works are shown in the upper table. Some authors have used ML algorithms and techniques like RF, LR, DT, ANN, and KNN to predict liver disease. However, the main difference is in its attributes. Attributes can greatly improve accuracy. As in most of the papers, accuracy is a maximum of 75%. On the other hand, we proposed a method of predicting Liver Disease from the early stage with corresponding risk factors. Our achieved highest accuracy is 86.14% which is quite good and it mainly defines the best way to fit using the dataset.

IX. LIMITATIONS AND FUTURE WORK

It performs essential and life-sustaining functions. A healthy person can perform any work perfectly, this study presents several limits that need to be addressed. First, we only gathered information from UCI Machine Learning Repository. But, multiple datasets and exterior validation might have much better performance and be more faithful. 2nd, we considered only 583 patients' data as a sample to evaluate the most significant risk factor. We also utilized k-fold cross-validation is dependable for small dataset and help reduce significant errors [2]. Finally, we utilized a classification method for automated ML variables integration, but learning this is certainly deep has been accustomed to improve better prediction. The proposed model is mainly a comparison-related model that has used different machine learning algorithms to find the perfect solutions and Random Forest performance is better with 86.14% accuracy. The risk factor has been used for Liver Disease symptoms probability in a human body. In the future, the proposed model will be implemented as a mobile application so that patients and doctors easily can find the leaves disease and risk factors of the symptoms.

X. CONCLUSION

Nowadays, liver disease is an incremental threat to mankind. People from all over the world are suffering and trying to combat, liver diseases. Synchronized with the advancement of medical technology, the rate of people affected by liver disease also increasing (Source data: World Health Organization 2018). Our proposed method works with 6 different algorithms LR, DT, MLP, RF, ANN, and KNN [16]. We measured the performance based on accuracy. So here, 72.89%, 81.32%, 60.24%, 86.14%, 75.61%, 65.52% accuracy given by LR, DT, MLP, RF, ANN, KNN. RF gives the highest accuracy among them and it defines the most acceptable fit for this dataset.

REFERENCES

[1] C. J. Pirola and S. Sookoian, "Multiomics biomarkers for the prediction of nonalcoholic fatty liver disease severity," *World J. Gastroenterol.*, vol. 24, no. 15, p. 1601, 2018.

[2] B. V. Ramana, M. S. P. Babu, and N. B. Venkateswarlu, "A critical comparative study of liver patients from USA and INDIA: an exploratory analysis," *Int. J. Comput. Sci. Issues*, vol. 9, no. 3, p. 506, 2012.

[3] M. A. Islam, S. Akter, M. S. Hossen, S. A. Keya, S. A. Tisha, and S. Hossain, "Risk Factor Prediction of Chronic Kidney Disease based on Machine Learning Algorithms," in *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, 2020, pp. 952–957.

[4] Y. M. Kadah, A. A. Farag, J. M. Zurada, A. M. Badawi, and A.-B. Youssef, "Classification algorithms for quantitative tissue characterization of diffuse liver disease from ultrasound images," *IEEE Trans. Med. Imaging*, vol. 15, no. 4, pp. 466–478, 1996.

[5] C.-L. Hsu *et al.*, "Role of fatty liver index and metabolic factors in the prediction of nonalcoholic fatty liver disease in a lean population receiving health checkup," *Clin. Transl. Gastroenterol.*, vol. 10, no. 5, 2019.

[6] M. A. Islam, M. S. Islam, M. S. Hossen, M. U. Emon, M. S. Keya, and A. Habib, "Machine Learning based Image Classification of Papaya Disease Recognition," in *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2020, pp. 1353–1360.

[7] M. S. Hossen, I. Haque, M. S. Islam, M. T. Ahmed, M. J. Nime, and M. A. Islam, "Deep Learning based Classification of Papaya Disease Recognition," in *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, 2020, pp. 945–951.

[8] V. Balasubramaniam, "Iot based biotelemetry for smart health care monitoring system," *J. Inf. Technol. Digit. World*, vol. 2, no. 3, pp. 183–190, 2020.

[9] S. Akter *et al.*, "Comprehensive Performance Assessment of Deep Learning Models in Early Prediction and Risk Identification of Chronic Kidney Disease," *IEEE Access*, vol. 9, pp. 165184–165206, 2021.

[10] A. Sathesh, "Computer Vision on IOT Based Patient Preference Management System," *J. Trends Comput. Sci. Smart Technol.*, vol.

2, no. 2, pp. 68–77, 2020.

[11] M. A. Islam *et al.*, “Forecast Breast Cancer Cells from Microscopic Biopsy Images using Big Transfer (BiT): A Deep Learning Approach.”

[12] V. Durai, S. Ramesh, and D. Kalthireddy, “Liver disease prediction using machine learning,” *Int. J. Adv. Res. Ideas Innov. Technol*, vol. 5, no. 2, pp. 1584–1588, 2019.

[13] J. Singh, S. Bagga, and R. Kaur, “Software-based prediction of liver disease with feature selection and classification techniques,” *Procedia Comput. Sci.*, vol. 167, pp. 1970–1980, 2020.

[14] J. Singh *et al.*, “Performance Assessment of Classification Algorithms on Early Detection of Liver Syndrome,” *Artif. Intell. Med.*, vol. 4, no. 1, pp. 42–46, 2018.

[15] A. S. Singh, M. Irfan, and A. Chowdhury, “Prediction of liver disease using classification algorithms,” in *2018 4th international conference on computing communication and automation (ICCCA)*, 2018, pp. 1–3.

[16] B. Khan, P. K. Shukla, M. K. Ahirwar, and M. Mishra, “Strategic analysis in prediction of liver disease using different classification algorithms,” in *Handbook of Research on Disease Prediction Through Data Analytics and Machine Learning*, IGI Global, 2021, pp. 437–449.

[17] A. K. M. S. Rahman, F. M. J. M. Shamrat, Z. Tasnim, J. Roy, and S. A. Hossain, “A Comparative Study On Liver Disease Prediction Using Supervised Machine Learning Algorithms,” *Int. J. Sci. Technol. Res.*, vol. 8, no. 11, pp. 419–422, 2019.

[18] I. Arshad, C. Dutta, T. Choudhury, and A. Thakral, “Liver disease detection due to excessive alcoholism using data mining techniques,” in *2018 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, 2018, pp. 163–168.

[19] T. A. Assegie, “Support Vector Machine And K-Nearest Neighbor Based Liver Disease Classification Model,” *Indones. J. Electron. Electromed. Eng. Med. informatics*, vol. 3, no. 1, pp. 9–14, 2021.

[20] R. A. Gani, “Survival analysis of hospitalized liver cirrhotic patients in Jakarta: 2 years follow up study,” *Indones. J. Gastroenterol. Hepatol. Dig. Endosc.*, vol. 22, no. 1, pp. 9–15, 2021.

[21] L. Rotty, N. Tendean, N. Lestari, and R. Adiwinata, “The Association between Interleukin-6 and Mean Platelet Volume Levels in Central Obesity with or without Non-Alcoholic Fatty Liver Disease,” *Indones. J. Gastroenterol. Hepatol. Dig. Endosc.*, vol. 21, no. 3, pp. 193–198, 2020.

[22] J. Tubung, I. K. Mariadi, and I. D. N. Wibawa, “Red Cell Distribution Width to Platelet Rasio is not inferior than Aspartate Aminotransferase to Platelet Ratio Indeks Score in Predicting Liver Fibrosis in Chronic Hepatitis B Patients at Sanglah General Hospital Denpasar,” *Indones. J. Gastroenterol. Hepatol. Dig. Endosc.*, vol. 19, no. 3, pp. 137–140, 2018.

[23] C.-C. Wu *et al.*, “Prediction of fatty liver disease using machine learning algorithms,” *Comput. Methods Programs Biomed.*, vol. 170, pp. 23–29, 2019.

[24] J. H. Joloudari, H. Saadatfar, A. Dehzangi, and S. Shamshirband, “Computer-aided decision-making for predicting liver disease using PSO-based optimized SVM with feature selection,” *Informatics Med. unlocked*, vol. 17, p. 100255, 2019.

[25] M. Kwak, E. Kim, E. J. Jang, and C. Lee, “The association of non-alcoholic fatty liver disease with lung function: A survey design analysis using propensity score,” *Respirology*, vol. 23, no. 1, pp. 82–88, 2018.

[26] S. R. Ghosh and S. Waheed, “Analysis of classification algorithms for liver disease diagnosis,” *J. Sci. Technol. Environ. Informatics*, vol. 5, no. 1, pp. 360–370, 2017.

[27] S. Vijayarani and S. Dhayanand, “Liver disease prediction using SVM and Naïve Bayes algorithms,” *Int. J. Sci. Eng. Technol. Res.*, vol. 4, no. 4, pp. 816–820, 2015.

[28] S. Dhamodharan, “Liver disease prediction using bayesian classification,” *COMPUSOFT An Int. J. Adv. Comput. Technol.*, 2014.

[29] E. M. Hashem and M. S. Mabrouk, “A study of support vector machine algorithm for liver disease diagnosis,” *Am. J. Intell. Syst.*, vol. 4, no. 1, pp. 9–14, 2014.

[30] R.-H. Lin, “An intelligent model for liver disease diagnosis,” *Artif. Intell. Med.*, vol. 47, no. 1, pp. 53–62, 2009.

[31] A. J. Dinu, R. Ganesan, F. Joseph, and V. Balaji, “A study on deep machine learning algorithms for diagnosis of diseases,” *Int. J. Appl. Eng. Res*, vol. 12, no. 17, pp. 6338–6346, 2017.

[32] A. S. Sulaiman *et al.*, “COVID-19 Related to Liver Impairment and Its Impact on Chronic Liver Disease,” *Indones. J. Gastroenterol. Hepatol. Dig. Endosc.*, vol. 21, no. 3, pp. 220–225, 2020.

[33] Z. T. Ghamari, F. Tadayon, and H. Mazdak, “Prevalence of liver cancer in Isfahan province, Iran,” *Indones. J. Cancer*, vol. 12, no. 2, pp. 56–59, 2018.

[34] E. Limantara, F. Kartawidjajaputra, and A. Suwanto, “Evaluation of potential gene expression as early markers of insulin resistance and non-alcoholic fatty liver disease in the Indonesian population,” *Indones. J. Biotechnol.*, vol. 23, no. 2, pp. 84–90, 2018.

[35] J. I. Z. Chen and P. Hengjinda, “Early prediction of coronary artery disease (cad) by machine learning method-a comparative study,” *J. Artif. Intell.*, vol. 3, no. 01, pp. 17–33, 2021.

[36] T. Vijayakumar, M. R. Vinothkanna, and M. Duraipandian, “Fusion based feature extraction analysis of ECG signal interpretation–a systematic approach,” *J. Artif. Intell.*, vol. 3, no. 01, pp. 1–16, 2021.

[37] M. A. Islam, R. Karim, F. Ahmed, M. S. Hossen, and S. Akter, “Broca’s Area of Brain to Analyze the Language Impairment Problem and Behavior Analysis of Autism,” in *Decision Intelligence Analytics and the Implementation of Strategic Business Management*, Springer, 2022, pp. 207–220.

[38] T. Khatun *et al.*, “Performance Analysis of Breast Cancer: A Machine Learning Approach,” in *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2021, pp. 1426–1434.

[39] L. Yogesh, M. Arunadevi, and C. P. S. Prakash, “Predicton of MRR & Surface Roughness in Wire EDM Machining using Decision Tree and Naive Bayes Algorithm,” in *2021 International Conference on Emerging Smart Computing and Informatics (ESCI)*, 2021, pp. 527–532.

[40] B. V. Ramana, M. S. P. Babu, and N. B. Venkateswarlu, “A critical study of selected classification algorithms for liver disease diagnosis,” *Int. J. Database Manag. Syst.*, vol. 3, no. 2, pp. 101–114, 2011.

[41] M. B. Priya, P. L. Juliet, and P. R. Tamilselvi, “Performance analysis of liver disease prediction using machine learning algorithms,” *Int. Res. J. Eng. Technol.*, vol. 5, no. 1, pp. 206–211, 2018.

[42] S. Sontakke, J. Lohokare, and R. Dani, “Diagnosis of liver diseases using machine learning,” in *2017 International Conference on Emerging Trends & Innovation in ICT (ICEI)*, 2017, pp. 129–133.