JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Liver Disease Prediction Using Machine Learning

Mrs. Veena Potdar[1], Mrs. Lavanya Santhosh[2], Kruthika Mannur[3], Rijuta D[4], Varshitha S[5], Sudhanva K S[6]

[1]Associate Professor, Department of Computer Science and Engineering, Dr. Ambedkar Institute of Technology, Bangalore, India

[2]Assistant Professor, Department of Computer Science and Engineering, Dr. Ambedkar Institute of Technology, Bangalore, India

[3,4,5,6]Student, Department of Computer Science and Engineering, Dr. Ambedkar Institute of Technology, Bangalore, India

Abstract - Since the inception of civilization, human beings have been the target of the effects of substances on the liver. The liver is one of the most important organs, being responsible for creating nutrients, metabolizing them and maintaining a non-toxic body condition. One of the most common types of liver disease is fatty liver disease. Fatty liver disease is one of the most dangerous human diseases and has a very serious impact on human life. Liver disease is also called by the name hepatic disease. It damages the liver. Some symptoms of liver disease may include weight loss and jaundice, and there are some different forms of liver disease. There are two different kinds of liver diseases. It is an alcoholic fatty liver disease that develops in people who consume a lot of alcohol. It results in damage to the liver and leads to cirrhosis. Non-alcohol fatty liver is one of the leading causes of liver disease worldwide. They accumulate fat in our liver. Human life can be saved by the detection of liver disease in its early stages. Data mining has become an easy method for liver disease prediction. The research being done to predict liver disease has a lot of challenging tasks. It is calculated from medical databases. To overcome this, research data mining techniques such as regression problems and classification are used. We can classify patients' risk levels by using data mining techniques. It predicts liver disease accurately and efficiently by using machine learning algorithms.

I. INTRODUCTION

The very first stage of damaged liver is an inflammation. Excessive fat in the liver is known as fatty liver. If we uncheck inflammation it leads to scarring. If there is scar tissue in the liver, it can lead to liver fibrosis over time. Gradually increasing of fat in hard scar tissue in the liver continuously can cause cirrhosis. The role of the liver is to ensure detoxification and metabolizes drugs in our body. Liver secrets bile. It extracts nutrients from carbohydrates, lipids, proteins. It stores vitamins and glycogen. It hoards energy in the form of sugar and provides for organism. Some important functions of liver are (i) Liver provides energy to our body by the help of storing vitamins, iron and sugar. (ii) It removes cholesterol from the body and controls the production of the same. (iii) Removing all waste products, such as drugs and other poisonous substances and clears the blood in the body. (iv) After cuts or injuries, it makes clotting factors to stop bleeding. (v) Removes bacteria from bloodstream to combat infection and produce immune factors in body. (vi) Liver secretes "bile" which helps in digesting food and absorbing nutrients. Removing all waste products, drugs and filtering harmful substance from the body. If liver doesn't function well, then it affects whole body. Once degradation begins, it damages the liver in a predictable manner. But in today's world liver disease has become the leading cause of death in the world. The main cause is liver disease & fatty liver in 377 Patients among 577 patients of the study [1]. A feature model has been implemented for better accuracy for disease prediction.

Proposed system concludes PSO feature selection for liver dataset. It predicts accuracy in four different phases:
i) In the first phase, liver patient datasets are collected from UCI Repository by applying min max normalization algorithm. (ii) In the second phase, the normalized liver patient datasets are extracted iii) In third phase, classification algorithm is applied. (iv) In fourth phase, accuracy is calculated [2].
By using machine learning approaches, we implement classification algorithm to identify liver patient based on performance factors. The graphical user interface will be developed by using python GUI [3]. Datamining techniques & algorithms are used to increase the accuracy and time. By using genetic algorithm accuracy increased to 93% [4]. To develop classification techniques to predict disease 11 algorithms were put it to datasets and compared with the terms of accuracy, precision and recall [5]. Research may be done in various decision tree techniques, predicted liver diseases. The various decision tree techniques such as RF, J48, LMT, RT, decision stump, REP Tree and Hoeffiding tree are used. The need for increasing accuracy furthermore has been done by using decision tree such as CART [6]. Liver disease have become leading cause of death. By using various algorithms such as decision tree, J48, ANN, Navie Bayes algorithms to classify the liver disease predictive or descriptive accuracy [7]. By using machine learning, it gives improved exactness on discovery of liver disease.[9].

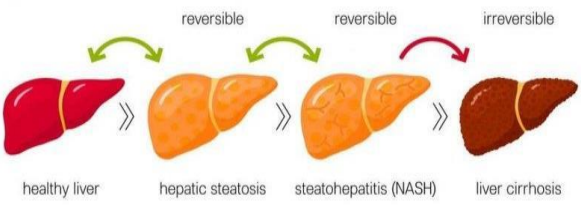

Fig. 1. Stages Of Liver Disease

II. EXISTING MODELS

1. [Software based Prediction of Liver Disease with Feature Selection and Classification Techniques Jagdeep Singha, Sachin Baggab, Ranjodh Kaur]

A web-based software application has been developed on basis of the software engineering life cycle model. There are four main phases: Planning and Analysis, Design and Build, Implementation, Validation, and delivery. Each phase contains several activities, and each phase is interconnected with other phases.

A. Preprocessing and Feature selection

To normalize the missing values, preprocessing techniques have been introduced. The missing values were replaced by null values along with their instances. Feature selection was followed to classify the appropriate attribute for classification. Using both filter and wrapper approaches, feature selection was carried out. The attributes with more than 70% correlation were initially excluded by correlation analysis from the dataset. The algorithm was implemented to estimate the value of different features in a dataset on the basis of random forests.

B. Planning and Analysis Phase

Planning phase includes the creation of ideas to support healthcare and technical team through the prediction of liver diseases. The main objective of planning phase is to plan the step involved in the development of prediction system using software engineering life cycle. In addition, challenging thing is to remove the gap between the software development members and health care specialists. In the analysis phase, the concern is to gather prediction system requirements and environmental considerations. The requirements involve the people from a different background area such as informaticists, physicians, patients etc.

C. Design and Build Phase

In design phase, the architecture model of liver diseases prediction software is established. The architecture defines user interface, segment, action and behaviour of the ILDP Software. The design document defines the technical plan to implement as per the requirements to build the system. The details of packages, programming language, platform, environment, and other technical/non-technical details are established.

D. Implementation Phase

In implementation phase, the development of ILDPS done as identified in the design phase. The main challenge in implementation phase is to implement the prediction system as per requirement, planning, and design. In the implementation phase, ILDPS is dealing with problems related to the performance, quality and debugging.

2. [REVIEW OF LIVER DISEASE PREDICTION USING MACHINE LEARNING ALGORITHM Vijay Panwar, Naved Choudhary, Sonam Mittal, Gaurav Sahu]

Data Collection: For this study, the Indian Liver Patient Dataset (ILPD) was selected from the UCI

Machine Learning repository. It is a sample of the whole Indian population taken from the area of Andhra Pradesh. There were 583 instances based on ten different biological parameters in the dataset. Based on these criteria, the class value was stated as either yes (416 cases) or no (167 cases), reflecting the liver.

B. Randomization and splitting of dataset

To build classification models, the features selected in the preceding phase were accepted. The dataset was initially randomized to produce an arbitrary sample permutation. Splitting of the dataset into training (70 percent of the dataset) and test (30 percent) sets was followed. The training set consisted of 389 cases and the evaluation set consisted of the remaining cases.

C. Classification algorithms

Classification algorithm is one of the greatest significant and applicable data mining techniques used to apply in disease prediction. Classification algorithm is the most common in several automatic medical health diagnoses.

Many of them show good classification accuracy. Different data mining algorithms like Naïve Bayes and Logistic Regression were implemented. The algorithms are briefly discussed below:

1. Naïve Bayes:

It is based on the Bayes theorem of conditional probability. The algorithm assumes that each attribute contributes to the total outcome independent of other attributes. In machine learning we are often interested in selecting the best hypothesis(h) given data(d). In a classification problem, our hypothesis(h) may be the class to assign for a new data instance(d). One of the easiest ways of selecting the most probable hypothesis given the data that we have & that we can use as our prior knowledge about the problem. Bayes' Theorem provides a way that we can calculate the probability of a hypothesis given our prior knowledge. Bayes' Theorem is stated as:

$$P(h|d) = (P(d|h) * P(h)) / P(d)$$

Where, P(h|d) is the probability of hypothesis h given the data d. This is called the posterior probability. P(d|h) is the probability of data d given that the hypothesis h was true. P(h) is the probability of hypothesis h being true (regardless of the data). This is called the prior probability of h. P(d) is the probability of the data (regardless of the hypothesis).

2. Logistic Regression:

Calculated Regression was for the most part utilized in natural research and applications in the mid-20th century. Logistic regression can deal with any number of numerical as well as absolute factors. In addition, it introduces a discrete parallelism somewhere in the range of 0 and 1. Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes. In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no). Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc. Logistic regression can be divided into following types Binary or Binomial. In such a kind of classification, a dependent variable will have only two possible types example, these variables may represent success or failure, yes or no, win or loss etc. Multinomial either 1 or 0. In such a kind of classification, dependent variable can have 3 or more possible unordered 0, for ordered types or the types having no quantitative significance. For example, these variables may represent "Type A" or "Type B" or "Type C". In such a kind of classification, dependent variable can have 3 or more possible ordered types or the types having a quantitative significance. For example, these variables may represent "poor" or "good", "very good", "Excellent" and each category can have the scores like 0,1,2,3.

3. [A Comparative Study on Liver Disease Prediction Using Supervised Machine Learning Algorithms, AKM Sazzadur Rahman, FM Javed Mehedi Shamrat, Zarrin Tasnim, Joy Roy, Syed Akhter Hussain]

Data Collection: In this experiment, we collect the dataset from the UCI Machine Learning Repository. In addition, the original dataset was collected from the northeast of Andhra Pradesh, India. This dataset consists of 583 liver patient's data whereas 75.64% male patients and 24.36% are female patients. This dataset has contained 11 particular parameters where we choose 10 parameters for our further analysis and 1 parameter as a target class. Such as

I. Age: Age of the patient
II. Gender: Gender of the Patients III.
III. TB: Total Bilirubin
IV. .DB: Direct Bilirubin

V. Alkphos: Alkaline Phosphotase
VI. Sgpt: Alamine Aminotransferase
VII. Sgot: Asparatate Aminotransferase
VIII. TP: Total Proteins
IX. ALB: Albumin
X. AG Ratio: Albumin and Globulin Ratio
XI. Selector field used to split the data into two sets (labeled by the experts)

A. Classification algorithms

1. Logistics Regression (LR)

Calculated Regression was for the most part utilized in natural research and applications in the mid-20th century. Logistic regression can deal with any number of numerical as well as absolute factors. In addition, it introduces a discrete parallel item somewhere in the range of 0 and 1. Strategic Regression processes the connection between the element factors by surveying probabilities (p) utilizing an underlying logistic function. Regression equation given as,

$$p = \frac{1}{1+e^{-(b_0+b_1x_1+\cdots+b_nx_n)}}$$

2.K Nearest Neighbors (KNN)

K nearest neighbors (KNN) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification predictive problems in industry. The following two properties would define KNN well-

Lazy learning algorithm − KNN is a lazy learning algorithm because it does not have a specialized training phase and uses all the data for training while classification.

Non-parametric learning algorithm − KNN is also a non-parametric learning algorithm because it doesn't assume anything about the underlying data.

2.Naïve Bayes:

It is based on the Bayes theorem of conditional probability. The algorithm assumes that each attribute contributes to the total outcome independent of other attributes [10]. In machine learning we are often interested in selecting the best hypothesis (h) given data(d). In a classification problem, our hypothesis(h) may be the class to assign for a new data instance (d). One of the easiest ways of selecting the most probable hypothesis given the data that we have that we can use as our prior knowledge about the problem. Bayes' Theorem provides a way that we can calculate the probability of a hypothesis given our prior knowledge. Bayes' Theorem is stated as:

P(h|d) = (P(d|h) * P(h)) / P(d)

Where, P(h|d) is the probability of hypothesis h given the data d. This is called the posterior probability. P(d|h) is the probability of data d given that the hypothesis h was true. P(h) is the probability of hypothesis h being true (regardless of the data). This is called the prior probability of h. P(d) is the probability of the data (regardless of the hypothesis).

4. [Statistical Machine Learning Approaches to Liver Disease Prediction Fahad Mostafa, Easin Hasan, Morgan Williamson, Hafiz Khan]

Data Description: Data was collected from the University of California Irvine Machine Learning Repository. The data set was included with laboratory reports of blood donors and non-blood donors with Hepatitis C and demographic information such as age and sex.

The response variable for classification was categorical variable: healthy individuals (i.e., blood donors) vs. patients with liver disease (i.e., non-blood donors) including its progress, e.g., hepatitis C, fibrosis, and cirrhosis. The data set contained 14 attributes such as ALB, ALP, BIL, choline esterase (CHE), GGT, AST, ALT, CREA, PROT, and cholesterol (CHOL). The sex and outcome variables were categorical, and the age variable was continuous. Hoffmann et al. [27] used machine learning algorithms to validate existing or to suggest potentially new decision trees using a subset of the same data set. The present study used 615 patients' data (376 males, 239 females).

A. Data Visualization and Target Labelling:

Missing data are quite a common scenario in the application of data science. In this study, data were investigated using different plots to detect groups of individuals who had liver disease and no liver disease. The target variable was modified into a binary category, labelled "0" for no liver disease and "1" for liver disease. The following method was used to fill out missing data for each predictor in the multivariate data. The missing values are needed to impute so that the data set remains in balance and to obtain a better estimation of prediction.

B. Multiple Imputation by Chained Equations for Missing Data:

Multiple imputation was used via the chained equations method to generate the missing data. For multivariate missing data, the R package [22] known as "MICE" was used for multiple imputations. This function auto detects certain variables with missing values. It basically uses predictive mean matching (PMM), which is a semiparametric imputation. It is very close to regression except missing items are randomly filled by regression prediction. The algorithms for MICE are given below.

Step 1: Start with imputing the mean. Mean imputations are considered "position holders";

Step 2: the "position holder" presents imputations for one variable ("Var") which are impeded to the missing items;

Step 3: "Var" is the response variable where the other variables are predictor variables in the linear regression model (under the same assumption);

Step 4: the missing values for "Var" are then replaced with imputed values from the regression model;

Step 5: Repeat steps 2–4 and produce the missing data. One iteration is needed foreach variable and, finally, the missing values. Ten such cycles were performed by Raghunathan et. Al .

I. Principal Component Analysis for Dimension Reduction

After selecting the features using PCA, the reduced data set was split into two parts, where 564 individuals were selected for the training data and 51 individuals were in the test data set. Supervised learning was carried out on the data set using ANN, RF, and SVM.

A variance importance ranking plot uses mean decrease accuracy and mean decrease Gini index to determine which variables are important. In order to describe the accuracy of a binary classification model, we often use the measures of precision sensitivity and specificity. Accuracy is the model's ability to correctly identify observations, while the precision measures the model's ability to distinguish between positive and negative observations. The sensitivity measures how many positive classifications are determined out of all the available positive classifications.

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

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

III. PROPOSED MODEL

1.ALGORITHM

To minimize bias (for example by overfitting), we randomly divided the dataset into training and testing 70% of training and 30% of testing.70% (n = 735) for feature selection and 30% (n = 314) for the model generation (see below).

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \times \sum_{j=1}^{m} |\beta$$

We used LASSO algorithm. It is one of the regression problem available to analyze the data. LASSO Regression is used for linear regression. Data point shrunk toward central point like mean. The model provides fewer parameters. LASSO encourage sparse model. LASSO Regression can reduce the slope to be exactly equal to zero. It uses predictors, that minimizes prediction error. It gives best accuracy. The cost function is given for LASSO is:

Cost(W) = RSS(W) + α (Sum of squares of weight)

We used python for this project. Python is high level programming language, general purpose, iterative, interpreted and oops. It reduces complexity & has fewer lines of codes. We have used powerful Lasso regression technique. It works magnitude of coefficients of feature with magnitude the error between predicted and actual observations. It is L1 regularization technique. It has minimized cost function and is given as Cost(W) = RSS(W) + α(Sum of squares of weight). There are three different cases for values of is α.

1. α = 0; it is a simple linear regression with same coefficient
2. α = ∞ All co efficient zero
3. 0 < α < ∞ co efficient between 0

The following code is used for training and prediction through Lasso regression.

ALGORITHM STEPS:

1.Lasso regression instantiates the value with alpha of 0.01.

2. The model fits for training the data.

3. Lasso regression predicts the training data.

4.To print MSE, MAE, RMSE, & R Squared on training dataset.

5. Repeat the steps on test dataset (Fig. 1).

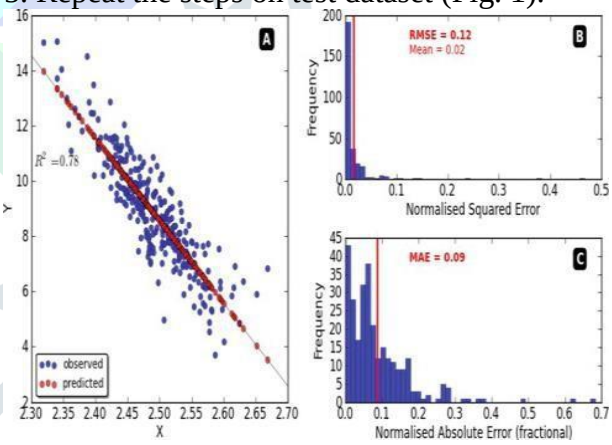

Fig.1 . Mean Absolute error or MSE

2. DATA SET

In this study, the liver patient dataset is chosen from UCI Repository based on the given liver attributes it predicts the disease. The dataset has 11 attributes & it predicts the liver disease and identify the patient result. The dataset is built on real number, categorical, integer types. The dataset consists of 11 characteristics, including age, gender, TB, DB, Alkphos, Sgpt, Sgot, TP, ALB, A/G ratio and class. The attributes age is real number age is one of the essential characteristics for growing the liver disease. Gender is greater risk factor for male 76% patients are affected for liver disease. It denoted by '1'and female 24% patients are affected for liver disease. It denoted by '0'. The

class attribute denotes “yes” with Boolean value “1” having liver disease. Class attribute denotes “no” with Boolean value “0” not having liver disease.

3. DATA PREPROCESSING:

1. Start is initial step to start the project.
2. Data base is collecting the data from external source like (Kaggle) before providing it to the model.
3. Data preprocessing which is used to transform data into information and efficient format with less effort.
4. Training is used to train an algorithm it minimizes the effects of data and later it sends to testing it test the model by making predictions.
5. We used Lasso algorithm to predict the liver disease it achieves with more effective and high accuracy.
6. Final step is predicting the disease. If the disease is predicting the output will be numerical “1” = yes, disease not predict value is “0” = no.

4. TOOLS AND LANGUAGE:

We use the Jupiter notebook as a tool and python 3.8 as a programming language. Some ML techniques are used for prediction of disease to get the result. Comparing various techniques such as decision tree, hybrid, lasso regression, NN, SVM, and RF finding the best technique to predict the disease.

ANACONDA: Anaconda installed new packages and tools. Install necessary libraries like matplotlib, pandas, NumPy, scikit learn, seaborn, pillow. Python main. py is used to execute the program (Fig. 2, 3, 4, 5, 6, 7, 8).

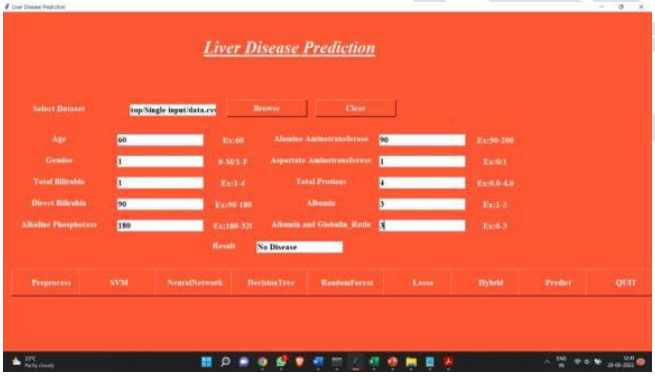

Fig.2. Implementation

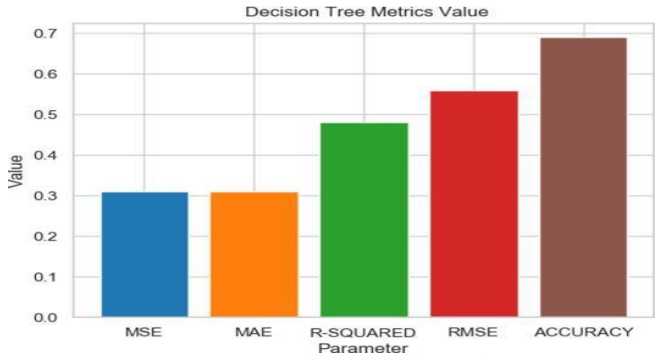

Fig.3. Decision tree metrics values

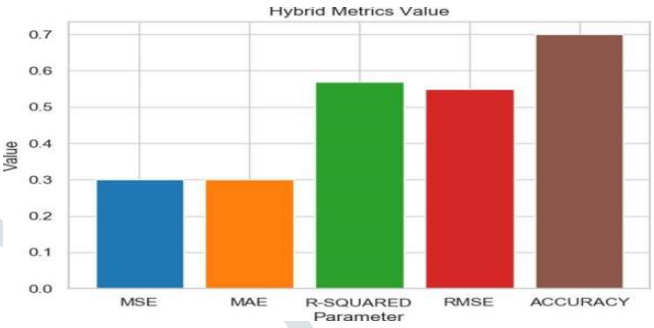

Fig.4. Hybrid metrics value

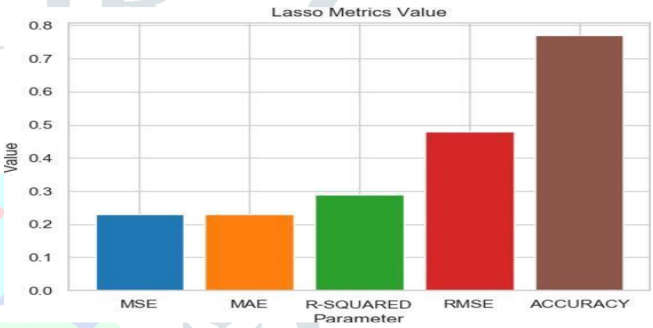

Fig .5. Lasso metrics value

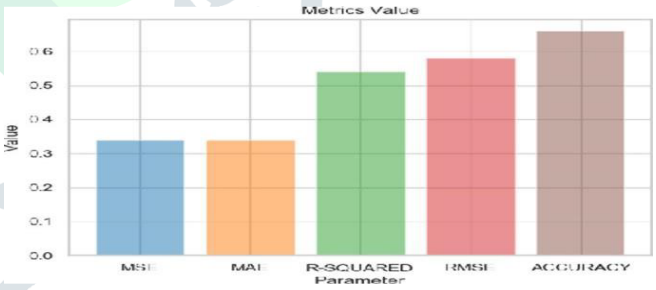

Fig. 6. Metrics value

5. MSE:

The mean squared error (MSE) compares the original image’s “real” pixel values to the degraded image. The MSE is the sum of the squares between the real and the noisy image of the “errors.” The error is the sum by which the original image values vary from those of the degraded image. It is portrayed as follows.

$$MSE^2 = \frac{1}{mn}\sum_{0}^{m-1}\sum_{0}^{n-1}\|f(i,j) - g(i,j) \quad (1)$$

$$MSE = (1/(m\ \ n))\ \ sum(sum((f - g).2)) \quad (2)$$

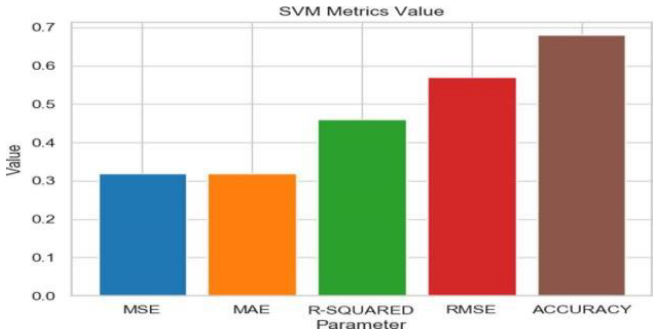

Fig. 7. SVM metrics value

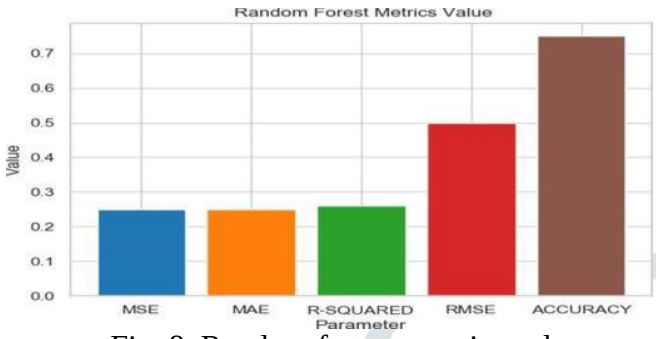

Fig. 8. Random forests metrics value

6. MAE:

It is the Difference between original and enhanced image is given as

$$MAE = |E(xN) - E(y)| \quad (3)$$

Where E(x)= average intensity of input image

E(y)= average intensity of enhanced image. The mean absolute error, where the estimation and the true value are is an average of the absolute errors. Note that alternative formulations can be weight variables that include relative frequencies. The mean absolute error was used on the same scale as the calculated data. This is regarded as a measure of scale dependent accuracy and can therefore not be used to use various scales to make comparisons between sequences.

R Squared Parameter:

It takes two measurements that exist in the population of detections, all classes or one as a subset of those detections, and then calculates a best fit line and R squared value for the resulting data point pairs.

RMSE:

The Root Mean Square Error (RMSE) is given as the MSE square root. This demonstrates that greater image quality is given by a higher PSNR and greater MSE & RMSE value

$$RMSE = \sqrt{MSE} \quad (4)$$

6. ACCURACY:

Accuracy can be defined as the percentage (TP +TN)/(TP + TN + FP + FN) of correctly classified instances. Where, respectively denote the sum of true positives, false negatives, false positives and true negatives (Fig. 9, 10, 11).

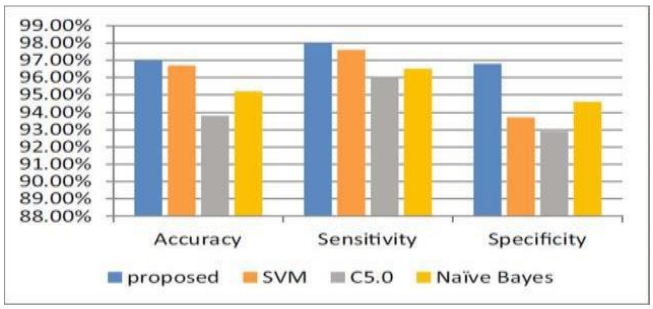

Fig. 9. The overall comparison between SVM, C5.0, and Naïve Bayes.

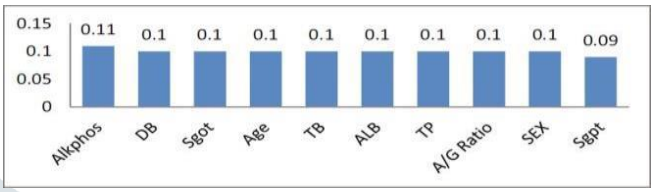

Fig. 10. Prediction of liver disease.

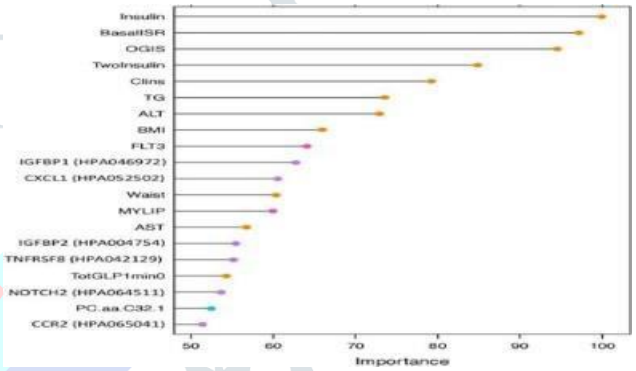

Fig. 11. Variable importance for the advanced model.

IV.RESULT

Image Segmentation Data Set

Dataset Characteristics:	Multivariate	Number of Instances:	2310	Area:	N/A
Attribute Characteristics:	Real	Number of Attributes:	19	Data Donated	1990-11-01
Associate Tasks:	Classification	Missing Values ?	No	Number of Web hits:	206228

V.CONCLUSION

The proposed system concludes that classification and Regression algorithms are applied for liver disease data set. Researchers have focused to save a human life and predict the liver disease in earlier stage. They used classification and regression algorithms such as decision tree, hybrid, lasso regression, neural network, SVM and random forest 69%, 70%, 77%, 66%, 68%, 75% to predict the liver disease. This algorithm gives better results comparing with other algorithms. However, in the future, we collect recent data for liver disease with advanced classification and regression techniques to predict the disease.

VI. REFERENCES

1. Wu, C.C., Yeh, W.C.: Prediction of fatty liver disease using machine learning algorithms. Comput. Methods Programs Biomed. 170, 23–29 (2019)

2. Banu Priya, M., Laura Juliet, P.: Performance analysis of liver disease prediction using machine learning algorithms. IRJET. 5(1) (2018)

3. Jacob, J., Chakkalakal Mathew, J.: Diagnosis of liver disease using machine learning techniques. IRJET. 5(4) (2018)

4. Hassoon, M.: Rule optimization of boosted C5.0 classification using genetic algorithm for liver disease prediction. IEEE (2017)

5. Bahramirad, S., Mustapha, A.: Classification of liver disease diagnosis: a comparative study. IEEE (2013)

6. Nahar, N., Ara, F.: Liver disease prediction by using different decision tree techniques. Int.

7. Baitharu, T.R., Pani, S.K.: Analysis of data mining techniques for healthcare decision support system using liver disorder dataset. Procedia Comput. Sci. 85, 862–870 (2016). https://doi.org/10.1016/j.procs.2016.05.276

8. Mehtaj Banu, H.: Liver disease prediction using machine learning algorithms. Int. J. Eng. Adv. Technol. 8(6), 1–3 (2019). https://doi.org/10.35940/ijeat.F8365.088619

9. Berzigotti, A., Ferraioli, G.: Novel ultrasound-based methods to assess liver disease: thegame has just begun. Dig. Liver Dis. 50(2), 107–112 (2018). https://doi.org/10.1016/j.dld.2017.11.019

10. Joloudari, J.H.: Computer aided decision making for predicting liver disease using PSO based optimized SVM with feature selection. Inform. Med. Unlocked 17, 100255 (2019)

11. Lin, R.H.: An intelligent model for liver disease diagnosis. Artif. Intell. Med. 47(1), 53–62 (2009)

12. Rahman, A.K.M.S. : A comparative study on liver disease prediction using supervised machine learning algorithms. Int. J. Sci. Technol. Res. 8(11), 419–422 (2019)