



Liver Disease Detection: A Review of Machine Learning Algorithms and Scope of Optimization

Aritra Pan, IMT Ghaziabad, India

 <https://orcid.org/0000-0001-5740-0804>

Shameek Mukhopadhyay, The Heritage Academy, India*

 <https://orcid.org/0000-0002-5764-2368>

Subrata Samanta, PricewaterhouseCoopers, India

ABSTRACT

In recent times, intelligent predictive systems are showing greater levels of accuracy and effectiveness in early detection of the critical diseases of cancer in the liver, lungs, etc. Predictive models assist medical practitioners to identify the diseases based on symptoms and health indicators like hormones, enzymes, age, blood counts, etc. This article focuses on proposing an optimal classification model to detect chronic liver disease by enhancing the prediction accuracy through cutting-edge analytics. The article proposes an enhanced framework on the original study by Ramana et al. It uses measures like precision and balanced accuracy to choose the most efficient classification algorithm in Indian and USA patient datasets using various factors like enzymes, age, etc. Using Youden's index, individual thresholds for each model were identified to increase the power of sensitivity and specificity, respectively. The study proposes a framework for highly accurate automated disease detection in the medical industry and helps in strategizing preventive measures for patients.

KEYWORDS

Balanced Accuracy, Classification Techniques, Liver Disease Detection, Precision, Youden's Index

1. INTRODUCTION

The largest internal organ of our body is the liver which weights around 3 pounds (Liet al., 2012). Liver performs different types of metabolic functions like filtering blood, producing bile, assisting in fat digestion, making proteins for blood clotting, metabolizing drugs, storing glucose, and most importantly detoxifying harmful chemicals which were discussed by Singhetal.(2017).Due to malfunctioning of liver, it may cause liver disease and may affect our health seriously. Liver disease

DOI: 10.4018/ijhisi.316666

*Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

are caused due to various reasons like consumption of contaminated food, inherited disorders, accumulation of excessive fat, damaged hepatocytes which is infected with bacteria, virus or fungi, and consumption of alcohol or drugs in excess as discussed by Linet al.(2010). An early diagnosis of liver disease may increase the patient's survival rate. To diagnose the liver disease along with various examination tests expert physicians are required, but it cannot assure the correct diagnosis was discussed by Takkaret. al (2017). Liver function tests majorly help in examining liver disorder. Key parameters in the test include albumin, alkaline phosphatase, total proteins, aspartate aminotransferase, alanine aminotransferase, direct bilirubin, total bilirubin, gamma-glutamyl transferase, prothrombin time, triglycerides platelet count and so on. Liver diseases are categorized into more than 100 type and the liver disease can be acute or chronic. Some liver disease has successful treatments while others don't have. Liver disease is generally caused by accumulation of excess fat, consumption of alcohol on long term basis, consumption of contaminated food, inherited disorders, and drug overdose.

Intelligent systems play a significant role in medical industry in terms of disease detection and prediction. Data mining algorithms, neural networks, statistical techniques are widely implemented on liver examination data to evaluate the sickness. Predictive modeling has been one of the broadly used intelligent techniques for automated detection of multiple diseases like cancer, cardiac arrhythmia, liver disease, lungs infection etc. There is a need of making predictive models more accurate for disease predictions like chronic liver disease, cancer, lungs disease, heart disease etc. Machine learning calculations play a significant role to the specialists in providing essential measurements, continuous information, and progressed examination about the patients' illness, lab test results, circulatory strain, clinical preliminary information, family history, and many more. Machine learning offers a guarantee for improving the detection and prediction of disease that has been made an interest in the biomedical field and also improve the decision-making process by increasing its objectivity. By way of machine learning techniques medical problems can be easily solved and thereby the diagnosis cost will be reduced. This study proposes an enhancement of the predictive models used in the original study by Ramana et al. (2011) to increase balanced accuracy and effectiveness of the models. The study focuses on two patient data sets (India and USA) and we have applied different machine learning techniques like Logistic Regression, K-Nearest Neighbors, Gradient Boosting Machine, Feedforward Neural Network, Support Vector Machine, C5.0, Naïve Bayes, Radio Frequency and the performance measures of these techniques were estimated on various perspectives such as Balanced Accuracy, Precision, recall, F1 – Score etc. to propose efficient classification algorithm for liver disease detection from various levels of enzymes, age and other factors and to predict the best classification algorithm in terms of accuracy, precision, specificity and sensitivity. Moreover, the performance was compared using the receiver operative characteristic (ROC). Classification algorithms are very effective and can be implemented in different automated medical diagnosis tools.

The aim of this study is to apply different machine learning techniques on liver datasets and thereby propose efficient classification algorithm. The remainder of this paper is structured in different sections where section 2 gives an idea of previous studies on liver disease diagnosis using classification algorithms. Section 3 describes the methodology and empirical design. Section 4 represents the finding of the study on liver datasets. Section 5 represents the implications of this study.

2. LITERATURE REVIEW

To accurately predict liver disease, the implementations of Machine learning techniques are done by different researchers from time to time which is as a result of improvement of technology and its effects in medical science. Arshad et al. (2018) discussed about detection of liver disease which are caused by excessive alcoholism using data mining techniques. They made a decision tree for the datasets and based on that the rules were generated. From UCI laboratory the data were collected and thereafter the training dataset was developed. Statistical techniques, data mining algorithms, neural networks have been widely deployed on liver examination data for evaluating the sickness. Based on the liver disorder a data classification technique

was proposed by Rajeswari et al. (2010). The training dataset is developed by collecting data from UCI repository which consists of 345 instances with 7 different attributes. This paper deals with results in the field of data classification obtained with Naïve Bayes algorithms. A proposal for identification of the liver disease among patients based on the 10 important attributes using a Decision Tree, Naive Bayes, and NB Tree algorithms was given by Alfisahrin et al. (2013) where they designed a model in WEKA tool. The result showed that NB Tree algorithm has the highest accuracy; however, the Naïve Bayes algorithm gives the fastest computation time. The concept of various classification techniques that assist the doctors to determine the disease quickly and efficiently was described by Jin et al. (2014). Various classifiers such as Naïve Bayes, Multi-Layer Perceptron, Decision Tree and k-NN were compared and analyzed based on several parameters like specificity, sensitivity and so on. Using Weka tool, the algorithms were implemented, and dataset was collected from UCI Repository. The experimental results showed that in terms of precision, Naïve Bayes gave the better classification results whereas Logistic Regression and Random Forest gave better results in terms of recall and sensitivity. Ramana et al. (2011) proposed the different types of liver dataset that is AP liver dataset and UCLA dataset and then evaluate the performance of the classification techniques from precision, accuracy, specificity, and sensitivity. The author said, AP liver dataset is better than the UCLA liver dataset. Using classification algorithm, they are support vector machine, C4.5, Back propagation neural network algorithm, and Naive Bayes classifier. Different classification algorithms were applied on ILPD-Indian Liver Patient Dataset and BUPA liver disorder dataset from UCI machine learning repository by Mazaheri et al. (2015). Durai et al. (2019) predicted liver diseases using machine learning in which they found that some of the machine learning approaches are not viable for a large volume of data and when it comes to the classification process and it is not necessary that the cohesion that a classifier shares with a particular set of data should stand viable for the rest of the training set. They identified that the quantity of data involved, features in the data, quality of data plays a major challenge for the accuracy of machine learning. Kuzhippallil et al. (2020) made a comparative analysis of machine learning techniques for Indian liver disease patients. They proposed method for building predictive model for liver disease using various supervised machine learning algorithms. They applied genetic algorithm in combination with XGBoost to fetch the best attributes required for prediction of liver disease and thereby different performance metrics were effectively utilized. Baitharu et al. (2016) presented an approach of diagnoses of liver disorder through an analysis of liver disorder datasets. The main focus of the research was to help the physicians with the medical decision-making process. Several algorithms were compared on various parameters such as Naive Bayes, Artificial Neural Network, ZeroRule algorithm (ZeroR), IBK, VFI, J48 and Multilayer Perceptron. The algorithms were implemented using the Weka tool and dataset was collected from UCI Repository. The experimental results showed that Multilayer Perceptron gave the better classification results as compared to other algorithms. Thus, Multilayer Perceptron can be further used to diagnose the liver disorder efficiently. Another concept for diagnoses of liver disease was proposed by Pathan et al. (2018). Various classification algorithms were used such as Naïve Bayes, Adaptive Boosting (Ada Boost), J48, Bagging and Random Forest. These algorithms were further compared based on the parameters such as Accuracy, Error rate and so on. In addition, pre-processing technique was utilized to divide the data into two groups- liver patients and non-liver patient that was accomplished using K means clustering algorithm. Further, the clustered dataset was applied to the various classification algorithms. The implementation of the different classification algorithms was performed using the Weka Tool. The overall comparison was done between Naïve Bayes, Ada Boost, J48, Bagging and Random Forest algorithms. After the comparison was performed, the comparative study showed that the Random Forest gave the better results as compared to the other algorithms. Vijayarani et al. (2015) demonstrated the predictive analysis of liver disorder using various classification algorithms. In this approach, Naïve Bayes and Support Vector Machine classification algorithms were used. These two algorithms were compared on the basis of performance parameters that include classification accuracy measures and execution time measures. Matlab tool was used to implement the proposed system and accordingly the dataset that had been collected from UCI Repository was evaluated. After the experimental results, it had been observed that Support Vector Machine outperformed Naïve Bayes Algorithm due to the highest classification

accuracy and can be used further in the prediction of liver disease. Singh et al. (2016) presented an efficient diagnostic system for the detection of liver disease that uses a unified two-phase approach based on PCA and K-nearest neighbor classifier (PCA-KNN). In terms of positive v. negative predictive value, accuracy, specificity and sensitivity the prediction model displayed exceptional results. Sug et al. (2012) in his study, the author compensates the insufficiency of liver disease disorder data usefully, he said a method-based oversampling in minor classes. Decision tree algorithm does not give high priority for minor classes for that reason using duplication in BUPA liver disease disorder dataset, increase the number of instances of minor class and proceed with two decision tree algorithms. Such that CART and C4.5 both algorithms are gives good result with oversampling for liver disease disorder dataset, but in future work can reduce the minor class increment to smaller percentage. Use of transient elastography for diagnosis of liver fibrosis and cirrhosis was carried out by Geng et al. (2016). It showed specificity of 88%, sensitivity of 81% and the area under the receiver operating characteristic (AUC - ROC) curve was 0.931.

Over the time there has been various research carried out to bring out better prediction models for liver disorder disease. It has been observed that through the usage of data mining algorithms various past researchers have focused on providing a better solution to the issue.

2.1 Research Gap and Objectives of the Study

In recent years, healthcare systems have started using modern and automated capabilities like machine learning, data mining techniques, artificial intelligence so as to improve the diagnosis and treatment in medical industry. This indeed creates a scope for providing excellent medical solutions to the patients. Healthcare Management is one of the areas which is using predictive analytics broadly for different objectives like disease detection, patient care, patient recovery, and drug formulation as discussed by Park et al. (2014). Healthcare management not only consists of hospitals and patient care, but it also incorporates Pharmaceuticals sector. Though liver disease is one of the common diseases present all over the world, it is still not easily detected at early stages, Lin et al. (2009); Faisal et al. (2018); Wu et al. (2017). Liver disease can be due to many factors like smoking, consumption of excessive alcohol, drinking arsenic contaminated water, obesity, low immunity and by inheritance. An early detection of this disease can help medical practitioners to make an informed decision and strategize treatment plan. Liver disease can be identified through analyzing different enzymes in blood Schiff et al. (2017). An efficient classification algorithm can help in early detection of liver disease in presence of necessary patient data. There are several studies Ramana et al. (2011); Sorich et al. (2003); Lin et al. (2009); Harper et al. (2005); Huang et al. (2009); Faisal et al.(2018); Wu et al. (2017) which have used different types of machine learning algorithms like Random Forest, Support Vector Machine (SVM), Neural Network, K nearest neighbor, Naïve Bayes and Decision Tree in prediction of chemical and medical datasets. This research is based on two datasets: Indian and USA from UCI datasets (ILPD (Indian Liver Patient Dataset) and Liver Disorders Data Set) repository to predict presence of liver disease. This work is an enhancement on the groundwork of Ramana et al. (2011) in terms of optimizing the model and enhancing the liver disease detection using broad set of parameters. In this research, we are proposing to enhance the power of predictive models in detecting Liver disease among patients. Our effort is to improve the machine learning model of the original work done by Ramana et al. (2011). The study used different health indicators likeage, enzyme details etc. This study will help medical industry to improvise and reduce errors in identification of diseases as well as accurate proposal for cure.

The above discussed literatures present an in-depth review of techniques involved in health care critical disease detection like liver disease. Literatures have outlined different machine learning algorithms as per their prediction power as well as broad set of factors as determinants of liver disease detection. Present article has framed following objectives for examination using 4-fold cross validation on two datasets (Indian and US patients):

- a) Predict and compare the accuracies of multiple machine learning models based on the original study by Ramana et al. (2011).

- b) Use of Youden's J Statistics to improve the classification models to receive best prediction results.
- c) Feature engineering to remove the high correlation among factors affecting liver disease detection accuracy.
- d) An optimized way of feature selection for best performing model.
- e) Use of grid search for model optimization.

The findings would be one of the primary analysis for understanding and predicting liver disease using robust machine learning model.

3. METHODOLOGY

This study focuses in identifying a patient with liver disease based on the selected attributes. We are proposing the classification model to be capable enough in identifying liver disease without any intervention from an expert doctor. This kind of model can be leveraged in medical industry, where it will help the medical practitioners to diagnose accurately within a short period of time.

3.1 Data Description

In this study, optimization of liver disease detection has been performed using different classification algorithms. Two Liver patient datasets, used in this study, are samples from Liver Disease (ILPD) India and Liver Disorders datasets of USA collected from University of California at Irvine (UCI) Machine Learning Repository. The attributes of Indian data set are age, gender, total bilirubin, direct bilirubin, alkaline phosphatase (alkphos), alamine aminotransferase (SGPT), aspartate aminotransferase (SGOT), total proteins, albumin and albumin and globulin Ratio (A/G ratio) are also provided in Table 1. The attributes of USA data set are mean corpuscular volume (mcv), alkaline phosphatase (alkphos), alanine aminotransferase (SGPT), aspartate aminotransferase (SGOT), gamma-glutamyltranspeptidase (gammagt), and number of half-pint equivalents of alcoholic beverages drunk per day (drinks) are also provided in Table 2. The common liver functional tests from both the data sets are alkphos, SGPT and SGOT. USA data set contains 345 patient records and INDIA data set contains 583 patient records. Figure 1 demonstrates the Label wise counts distributions with disease vs without disease ratio of 416:167 and 145:200 for INDIA and USA

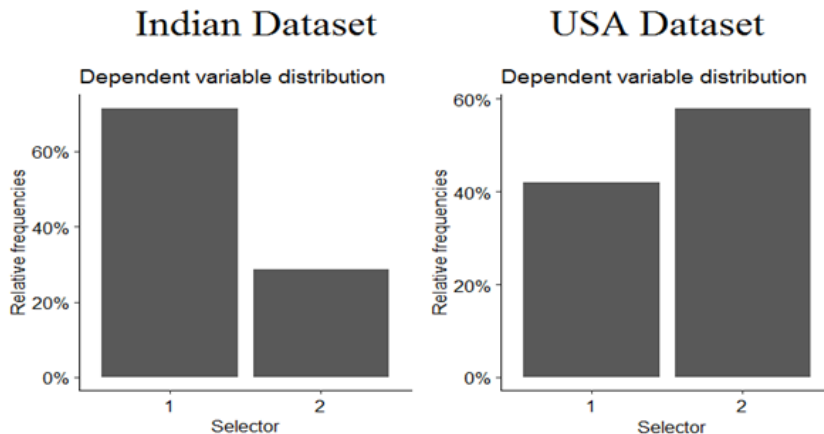
Table 1.
Data description of Indian liver patient's dataset

INDIAN DATA		
	Mean	Std. Dev.
Age	44.75	16.18983
total bilirubin (TB)	3.299	6.209522
direct bilirubin (DB)	1.486	2.808498
alkaline phosphatase (alkphos)	290.6	242.938
alanine aminotransferase (SGPT)	80.71	182.6204
aspartate aminotransferase (SGOT)	109.9	288.9185
total proteins (TP)	6.483	1.085451
albumin (ALB)	3.142	0.795519
albumin and globulin Ratio (A.G.Ratio)	0.9471	0.3195921

Table 2.
Data description of USA liver patient's dataset

USA DATA		
	Mean	Std. Dev.
mean corpuscular volume (mcv)	90.16	4.448096
alkaline phosphatase (alkphos)	69.87	18.34767
alanine aminotransferase (sgpt)	30.41	19.51231
aspartate aminotransferase (sgot)	24.64	10.06449
gamma-glutamyltranspeptidase (gammagt)	38.28	39.25462
alcoholic beverages drunk per day (drinks)	3.455	3.337835

Figure 1.
Relative frequencies of dependent variable (1 represents patients with liver disease & 2 represents patients without liver disease)



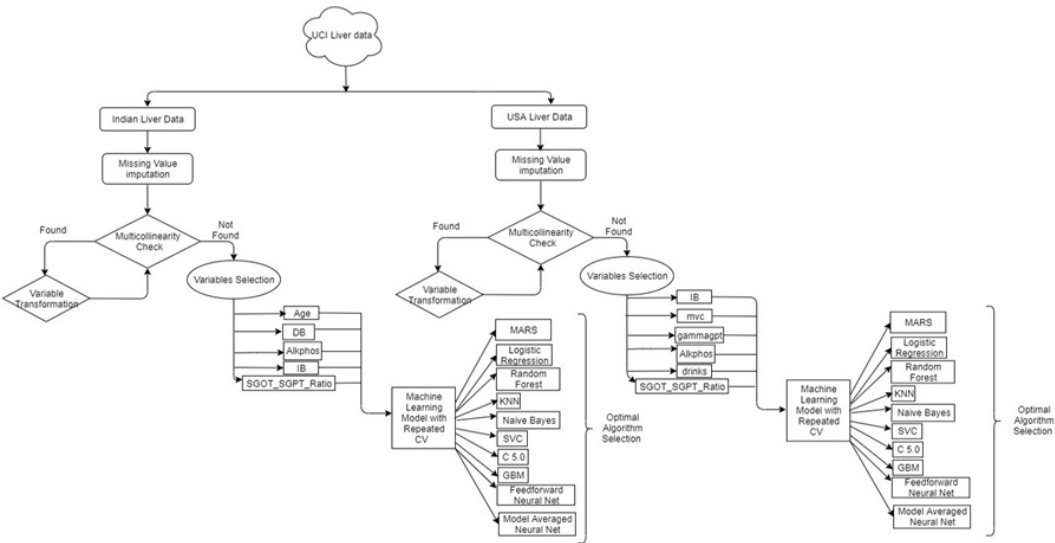
dataset respectively. To understand this, Figure 1 is showing Relative frequencies of dependent variable for Indian and USA dataset.

3.2 Proposed Framework for Study

For accomplishing this task of classification problem, we have followed several steps including data preprocessing, training machine learning model, tuning etc. to reach a conclusion. Firstly, Indian/USA liver patient data have been collected from UCI/UCLA repository. Then exploratory data analysis has performed to investigate the presence of missing values and multi collinearity in the data. For the missing values in the selected column (A.G. Ratio), missing value imputation have performed to get completeness of data. Besides, to avoid the multi collinearity in the data, checking has been incorporated for Indian data, in this paper we carry out variable transformation to counter the effect of it.

Similarly, the process follows all the steps including missing value handling, multi collinearity check and feature engineering for USA data before training the machine learning models. After that variable selection process have been performed based on the relative importance of the variables in presence of dependent variable, so that our model would be easier to understand and interpretable. After all these steps, now the data is ready for machine learning algorithms to be applied on it. We have performed repeated cross validation while training different machine

Figure 2. Process Flow



learning algorithms on training data set. Then we analyzed the F1 scores of all machine learning algorithms to select the optimal classifier. To understand this, **Figure 2** is showing flowchart of entire process have been mentioned.

3.2.1 Missing Value Handling

The Indian/USA dataset faced with the issue of missing values and mentioned in **Table 3 & 4**. This research uses random forest algorithm to fill up the missing values (Young, 2017). Mice package (Buuren et al., 2011) in R has enabled with the option of missing value imputation using multiple algorithms like random forest, predictive mean matching, weighted predictive mean matching etc.

Random forest is one of the imputation methods built on the Mice framework (Shah et al., 2014). As the missing values were continuous variables, Mice with random forests assigns the missing

Table 3. Missing value counts of Indian liver patient's dataset

INDIAN DATA	
	Missing value
Age	0%
TB	0%
DB	0%
Alkphos	0%
Sgpt	0%
Sgot	0%
TP	0%
ALB	0%
A.G.Ratio	0.69%

Table 4. Missing value counts of USA liver patient's dataset

USA DATA	
	Missing value
mcv	0%
alkphos	0%
sgpt	0%
sgot	0%
gammagt	0%
drinks	0%

values by applying random produces from independent normal distributions which are centered on the means predicted from random forests. Out-of-bag mean square of error is used as estimator of the residual variance (Young, 2017).

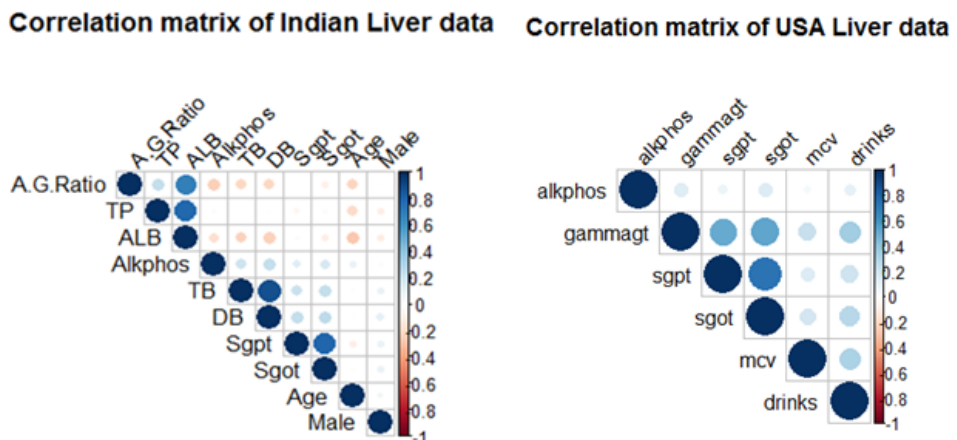
3.2.2 Feature Engineering & Correlation Analysis

This research finds high correlation (> 0.6) between variables SGOT and SGPT in both the datasets (**Figure 3**). We have also found DB and TB to be highly correlated (> 0.6) in Indian dataset. To mitigate the issue of high correlation, we have created two new variables. In USA dataset, we have introduced a ratio between SGOT and SGPT which is a significant parameter of identifying liver disease (Cohen et al., 1979). For the other two correlated variables DB and TB, we have computed IB or Indirect Bilirubin by subtracting DB (direct bilirubin) from TB (total bilirubin) which is practiced by pharmacists (Tietze, 2011). These features additions removed the correlation from data after removal of TB, SGOT and SGPT from their respective datasets.

3.2.3 Feature Selection

To extract useful information from the high volumes of data, one must use statistical techniques to reduce the noise or redundant data. This is because one often need not use every feature at one's disposal to train a model. One can improve his model by feeding in only those features that are uncorrelated and non-redundant. This is where feature selection plays an important

Figure 3. Correlation analysis of Indian and USA dataset



role. Not only it helps in training the model faster but also reduces the complexity of the model, makes it easier to interpret and improves the accuracy, precision or recall, whatever may the performance metric be.

We have used Boruta algorithm (Kursa et al., 2010) where the shadow features get created. The Boruta algorithm is a wrapper built around the random forest classification algorithm. It tries to capture all the important, interesting features one might have in his dataset with respect to an outcome variable. Using Boruta algorithm we have duplicated the dataset and shuffled the values in each column. Then we run random forest classifiers on the merged dataset to estimate the variable importance of each feature based on mean decrease accuracy measure. Maximum Z score (MZSA) among shadow variables have been calculated and tag the variables as ‘important’ (**Figure 4 and 5**) if they have notably higher importance than MZSA. Until all features are either tagged with one category, we repeat the same steps for predefined number of iterations.

For both the Indian and US liver data the missing values were imputed, and correlation was checked. If found the variable is transformed and when not found the variable selection is done on the ROC curve using Boruta algorithm.

3.2.4 Machine Learning Algorithms

After performing all the above procedures, we finalize our model formulation for model training.

Indian dataset:

$$\text{Selector} \sim \text{Age} + \text{DB} + \text{Alkphos} + \text{IB} + \text{SGOT_SGPT_Ratio} \quad (1a)$$

USA dataset:

$$\text{Selector} \sim \text{mcv} + \text{Alkphos} + \text{gammagt} + \text{drinks} + \text{SGOT_SGPT_Ratio} \quad (1b)$$

3.2.4.1 Logistic Regression:

Figure 4. Feature selection of Indian liver patient dataset

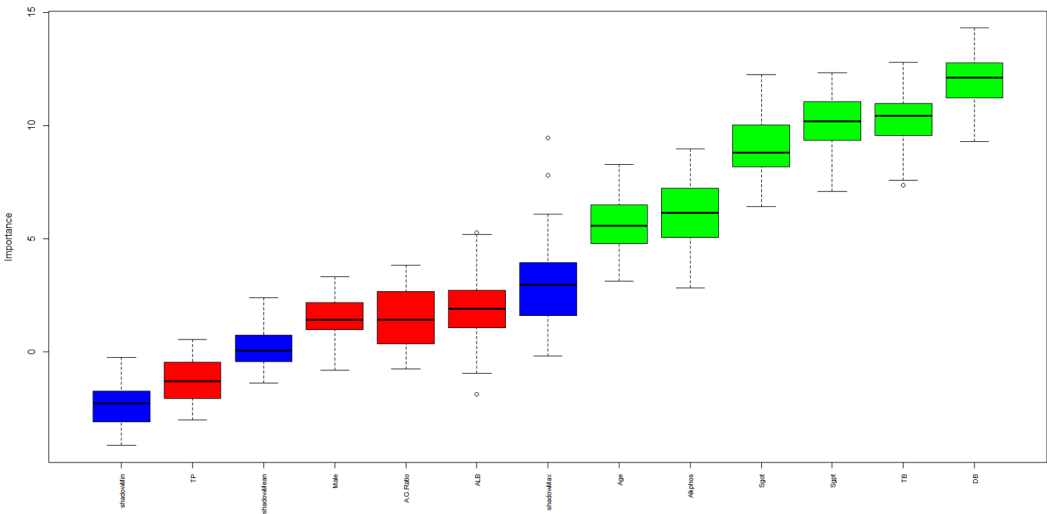
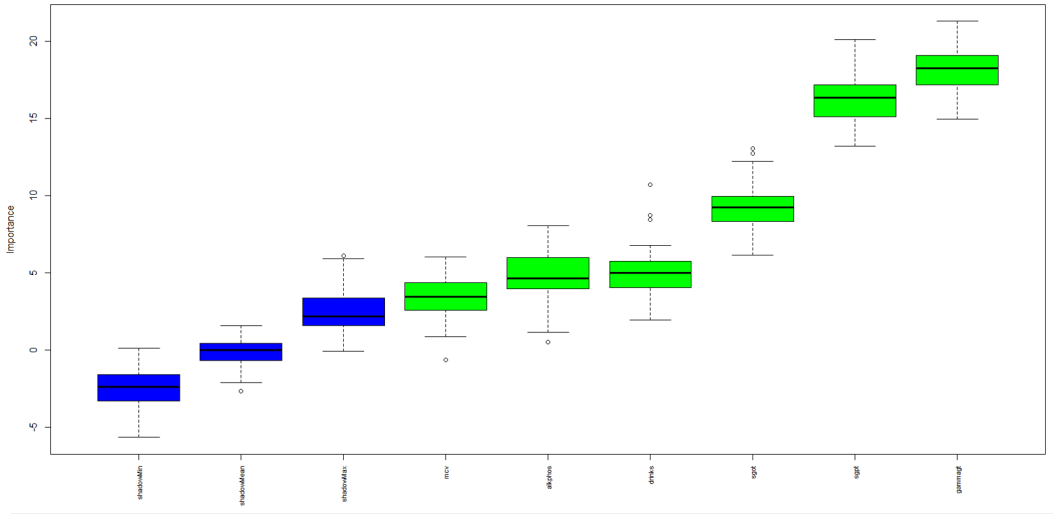


Figure 5. Feature selection of USA liver patient dataset



Logistic regression is a well-known statistical method to examine problems with binary outcome variables. Let presume two outcomes of the dependent variable are 1 and 0, respectively, the mathematical representation of logistic regression model (Kleinbaum et al., 2002) is shown below (Eq. 2a)-

$$\log it(Y = 1) = \ln \left[\frac{P(Y = 1)}{P(Y = 0)} \right] = \alpha + \beta X \quad (2a)$$

where Y is the dependent variable, 1 is the predicted outcome; X is the independent variable vector; α and β are parameters that will be identified by maximum likelihood estimation using the training data.

The probability $P(Y = 1)$ of a test case can be calculated using (Eq. 2b) and compared with the predefined threshold such that the class label of the test case can be determined.

$$P(Y = 1) = \frac{1}{1 + e^{-(\alpha + \beta X)}} \quad (2b)$$

3.2.4.2 Naive Bayes

The main goal of a Classifier is to accurately classify the class with respect to each instance in a given data set. This Classifier is based on Bayes' Theorem, in which supervised classification technique has been used to predict the class from the attributes of a data set. Bayesian Classifier is stated as Rish (2001) (Eq. 3) –

$$P(X) = \frac{P(X | C)P(C)}{P(X)} \quad (3)$$

Where, the posterior probability, $P(X)$ (the probability of an attribute value X belonging to a class C), is calculated using Class Prior Probability $P(C)$ (probability of class), Predictor Prior Probability $P(X)$ (probability of attribute value) and Likelihood $P(X | C)$ (probability of attribute value X given class C).

3.2.4.3 K Nearest Neighbor

The KNN classifier approximate a function $f: x \rightarrow y$ such as $x = \{x_{i1}, x_{i2}, \dots, x_{in}, y_i\}$ under some distance measures, finding k samples whose distances to the test sample are the smallest among all the training samples, and then predict the class of the test sample by taking the mode class of the k training samples. The distance between two samples is calculated as follows (Farahnakian et al., 2013) (Eq. 4).

$$d(x_i, x_j) = \left(\sum_{h=1}^m (x_{ih} - x_{jh})^2 \right)^{1/2} \quad (4)$$

For a test sample z , we calculate the distances $d(z, x_i)$ between it and each sample in the training data set $x = \{x_i, i = 1 \times \times \times, n\}$, and find m training samples $\{x_{NB(i)}, i = 1 \dots, k\}$ whose distances are the smallest. Based on the label values $\{y_{NB(i)}, i = 1 \dots, m\}$ corresponding to these k training samples, we can calculate the label estimate \hat{y}_z for the test sample z .

We used mean method for calculating the label estimate (\hat{y}_z) in this paper (Eq. 5).

$$\hat{y}_z = \frac{1}{k} \sum_{i=1}^k y_{NB(i)} \quad (5)$$

3.2.4.4 Support Vector Machine

The idea of support vector machine is to find the optimal hyperplane which maximize the margin between two classes. Mathematically, the optimal values for the hyperplane parameters w (e.g. w_0) and b (e.g. b_0). This algorithm finds the optimal separating hyperplane in such a way that $w_0 \cdot x + b_0 = 0$, then for predicting an unseen pattern, x_i , can be classified by the decision rule $f(x) = \text{sign}(w_0 \cdot x + b_0)$. Each x_i , has a corresponding value y_i , where $y_i \in \{1, -1\}$, while w and b are parameters of the hyperplane. The nearest data points to the maximum margin hyperplane lie on the planes as Furey et al. (2000) in (Eq. 6, Eq. 7):

$$(W \cdot X) + b = +1 \text{ for } y = +1 \quad (6)$$

$$(W \cdot X) + b = -1 \text{ for } y = -1 \quad (7)$$

By rescaling w and b , with no loss of generality, and grouping the above constraints in a single formula (Eq. 8), where $y = +1$ for class w_1 and $y = -1$ for class w_2 .

$$y_i f(x_i) \geq 1 \quad (8)$$

3.2.4.5 Random Forest

Random forest is an ensemble method where the predictions of several trees are combined to make a prediction. Whole dataset has been sampled randomly and each tree predictor is trained on a specific sample of data.

If N is the number of samples for constructing a tree, and M is the number of variables. Each tree will be constructed based on lowest classification error among finite number of predictors on a particular node in the following two steps (Wang et al., 2015):

1. Randomly select K samples from the whole dataset and divide the samples into training set and testing set;
2. At each tree node, randomly choose m variables and select the best split based on these variables to split the node.

Finally, taking voting of the K sample results as the final prediction result. After K 's round of training, you can get a classification prediction model sequence $\{h(X; ,_1), h(X; ,_2), \dots, h(X; ,_k)\}$. And the final prediction result is as follows (Eq. 9):

$$\bar{h}(X) = \text{mode} \sum_{k=1}^K h(X; ,_k) \quad (9)$$

3.2.4.6 Gradient Boosting

Gradient Boosting is an ensemble learning algorithm that constructs a greedy set of decision trees at training time and makes a class prediction by taking a majority vote. The model is an ensemble of k trees (Gupta et al., 2016), formalized as (Eq. 10):

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), f_k \in F \quad (10)$$

The model is built in a greedy manner, with an aim to minimize the objective function as (Eq. 11):

$$Obj = \sum_i^n l(\hat{y}_i, y) + \sum_k \Omega(f_k) \quad (11)$$

To find the function $f_t(x)$, the following objective function is minimized at round t (Eq. 12)

$$Obj = \sum_{i=1}^n l(y, \hat{y}_i^{(t)}) + \sum_{i=1}^n \Omega(f_k) \quad (12)$$

A quadratic approximation allows us to define the objective function at round t as (Eq. 13)

$$Obj^{(t)} = \sum_{i=1}^n \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_k) \quad (13)$$

Thus, the search for the functions f_t depend on the objective only via the first and second gradients of the loss functions, which are g_i and h_i respectively.

3.2.4.7 Multivariate Adaptive Regression Spline (MARS)

MARS is a regression analysis method proposed by Friedman, which is a nonparametric regression method. The complex nonlinear relationship between the variables is pretended by the spline function. MARS model has strong adaptability and model interpretation ability. The MARS model can be expressed (Onak et al., 2017) as (Eq. 14):

$$y = c_0 + \sum_{m=1}^M c_m B_m(x) \quad (14)$$

Where y is a dependent variable predicted by MARS, c_0 is a constant. The is an m -order basis function and c_m being the coefficient of the m -order basis function.

The basis function is in the form of $\max(0, x - c)$, $\max(0, c - x)$ or the product of the two. In the expression, c is a constant called knot. MARS always converges to the same set of basis functions with the same initial data set. Thus, it automatically simulates the nonlinearity and interaction as the weighted sum of basis functions.

3.2.4.8 Decision Tree

Decision tree is constructed by repeated splits of subsets into two descendant subsets. Every split has a decision-making mechanism for the input variables, and the answers of “yes” and “no” lead respectively to each descendant subset. Every subset is called a “node”. A leaf node is a node without further splits, and has an output value and a rule, which can be expressed in the form of “if ...then...”. When the decision tree is applied to new data, for which the class is unknown, it gives a prediction of the class (AnujaPriyama et al., 2013). This algorithm uses different impurity measures like Gini-index, information gain and some distance-based measures to choose an input feature to be associated with an internal node.

3.2.4.9 Neural Network

3.2.4.9.1 Feed Forward Neural Network (FNN)

A feed forward neural network consists of input layer having N_i nodes; the hidden layer having N_h hidden nodes; output layer having N_o output nodes. In this paper, the hidden transfer function and the output transfer function are both sigmoid functions. The computed output of the i -th node in the output layer is defined (Fitriyah et al., 2018) as follows (Eq. 15):

$$y_i = f \left(\sum_{j=1}^{N_h} w_{ij} f \left(\sum_{k=1}^{N_i} v_{jk} x_k + \theta_{vj} \right) + \theta_{wi} \right), i = 1, \dots, N_o \quad (15)$$

where y_i is the output of the i th node in the output layer; x_k is the input of the k th node in the input layer; w_{ij} is the connective weight between nodes in the hidden and output layers; v_{jk} represents the connective weight between the nodes in the input and hidden layers; and θ_{wi} (θ_{vj}) are bias terms that represent the threshold of the transfer function f . The learning error E can be calculated by the following formulation f .

The learning error E can be calculated by the following formulation (Eq. 16):

$$E = \sum_{k=1}^q \frac{E_k}{q \cdot O} \text{ where } E_k = \sum_{i=1}^O (y_i^k - C_i^k)^2 \quad (16)$$

where q is the number of total training samples, $(y_i^k - C_i^k)$ is the error of the actual output and desired output of the i th output unit when the k th training sample is used for training.

3.2.4.9.2 Model Averaged Neural Network

Model averaged neural network is a model where the same neural network model is fit using different random number seeds and then taking an ensemble approach for prediction (Kuhn, 2008). For classification problem, the model scores are first averaged, then translated to predicted classes. The number of models fit is controlled by the argument repeats which is passed down to the model in caret using repeats option. Generally model fitting and aggregation is performed by bootstrap aggregation, which is optimized to provide better predictive performance if the number of models is high enough.

3.2.5 Performance Evaluation of Machine Learning Algorithm

This research uses classification models like logistic regression, naïve bays, kNN, SVC, random forest, gradient boosting machine, C5.0 and feedforward neural network using 4-Fold cross validation and grid search parameters. Grid search for wide range of parameters have been used to gain maximum output from the models. However, given the size of the datasets, we were limited in the range of values for each parameter to avoid overfitting. For each of these models we have set an initial cut-off value of 0.5. For the cut-off value, we have estimated true positive, true negative, false positive and false negative values along with Kappa from confusion matrix (Stehman, 1997). We have computed different parameters viz. Accuracy, Sensitivity, Specificity, Precision, Negative Predictive Value, Miss Rate, Fall-Out, False Discovery Rate, False Omission Rate, Threat Score, Balanced Accuracy, Informedness, and F1 Score.

Balanced Accuracy: Balanced accuracy is calculated as the mean of the proportion of correctly classified liver and non-liver patients points of each class individually.

Informedness: Informedness measures how our model is informed about positives and negatives predictions by considering both real positives(RP) and real negatives(RN).

Threat Score (TS): The Threat Score (TS) or Critical Success Index (CSI) combines the fraction of predicted liver disease that is forecast correctly and the fraction of “yes” forecasts that were wrong into one score for low frequency events.

False Discovery Rate: The false discovery rate (FDR) is the expected proportion of positive. In our case the false discovery rate is when model predicted someone as liver patient, but he doesn't actually have the disease.

False Omission Rate: False omission rate measures the proportion of false negatives which are incorrectly rejected. In our case this indicate the number of wrongly rejected predictions termed as non-liver patients but when it is actually a liver patient.

After achieving the values using default classification threshold of 0.5, we have tuned the model using Youden's J Statistics or Youden's Index, Youden (1950) to achieve maximum Sensitivity and Specificity to maximize detection accuracy. This method helped the models to gain maximum Informedness in terms of probability of an informed decision.

4. FINDINGS OF THE STUDY

The primary objective of this study is to accurately identify the presence of liver disease in a patient. Liver disease datasets of India and USA are taken from UCI (University of California at Irvine)

machine learning repository. These datasets reflect presence or absence of liver disorder in patients using the various health examinations results of medical tests performed on patients. Key features from the tests for Indian patients include age, total bilirubin, direct bilirubin, albumin and globulin ratio, alkaline phosphatase, albumin, alanine aminotransferase, aspartate aminotransferase and total proteins. For the USA dataset, key features are mean corpuscular volume, alkaline phosphatase, alanine aminotransferase, aspartate aminotransferase, gamma-glutamyltranseptidase and alcoholic beverages drunk per day. Indian Liver patient dataset contains 583 samples belonging to two distinct classes (416 or 71.35% are cases of patient with liver disorder and 167 or 28.65% are of healthy individuals). For USA the sample size is 345 out of which 42% are cases of patient with liver disorder.

The classification methods implemented for the study were Logistic Regression, Random Forest, K-Nearest Neighbor (KNN), Naïve Bayes, Support Vector Machine (SVM), Decision Tree C 5.0, Gradient Boosting Machine (GBM), Multivariate Adaptive Regression Spline (MARS) and Neural Network (Feedforward and Model Average) based approaches. These approaches include 4-fold cross validation technique and optimization of the models with Youden's index. It was observed from the experimental results in Table 5 and Table 6, that all of the algorithms shown significant prediction performances improvements after tuning with Youden's index. The study considers Balanced Accuracy as a better evaluation metrics compared to Accuracy. Although, Random Forest which was tuned for optimal number of trees based on minimal OBB error, showed best accuracy rates with optimized cut-off value using Youden's Index for both India and USA liver datasets. We have taken a range of 5 to 2000 number of trees to grow the random forest to select the best random forest model. Model Averaged Neural Network showed significant improvement on results in USA dataset which was designed using 5 input, 1 hidden and one output layer in its structure.

Prior to the machine learning models, the study focuses on data engineering techniques which were applied on both India and USA liver datasets. The missing values found in India dataset were imputed based on predicted values using random forest technique. This eliminated risk of losing information by removing records with missing values. This study also checked for high correlation among explanatory variables. To deal with this issue, the study formulated new explanatory variables by doing feature engineering on highly correlated independent variables. Formulation of direct DB, indirect DB and SGOT – SGPT ratio were at par with medical literatures (Cohen et al., 1979; Tietze, 2011). This is one of the unique approaches followed by this study which was not present in the base study by Ramana et al. (2011) as well as other literature (Takkare et al., 2017; Singh et al., 2016).

This study focuses on broad set of evaluation metrics in identification of optimal classification algorithm compared to previous studies made on liver disease identification (Ramana et al., 2011), (Takkare et al., 2017), chronic disease classification (Jain et al., 2020), heart disease prediction (Priyanga et al., 2018). Using Youden's Index, this research was able to increase the precision of the models significantly compared to values against default threshold of 0.5. Youden's Index helped the models to increase probability of an informed decision which reduces negative predictive value and False Positive Rate (Table 5 and 6). False positives were significantly reduced which only increase Precisions in the models. In Indian data, almost all the model's accuracy took a dip except Random Forest and Support Vector Classification where accuracy increased after use of cut-off value from Youden's Index. Random Forest surpassed all other algorithms in terms of performance on both the datasets. Informedness saw significant jump in SVC compared to all other models in Indian liver dataset. The approach of using Youden's index helped in significant improvement over Precision which also ensures effectiveness in identifying a patient with liver disease. However, this is found for the Indian dataset only. For USA dataset, the study sees an improvement in Negative Predictive Value which helps in identifying healthy patients by reducing number of false negatives in identification. Though, overall balanced accuracy improved significantly for both datasets after use of Youden's index. Apart from the random forest, C 5.0 and GBM showed the significant accuracy in predictions. Informedness also took a jump compared in all the models with Youden's index compared to default threshold value 0.5.

Table 5. Machine Learning Models Evaluation Metrics of Indian Dataset

Models using 4-Fold Cross Validation	Logistic Regression	Logistic Regression	Naive Bayes	Naive Bayes	kNN	kNN	SVC	SVC	Random Forest (optimal ntree = 200)	Random Forest (optimal ntree = 200)	Gradient Boosting Machine	Gradient Boosting Machine	C 5.0	C 5.0	Feedforward Neural Network	Feedforward Neural Network	Model Averaged Neural Network	Model Averaged Neural Network	Multivariate Adaptive Regression Spline	Multivariate Adaptive Regression Spline
Cut-Off	0.500	0.776	0.500	0.640	0.500	0.678	0.500	0.736	0.500	0.600	0.500	0.724	0.500	0.706	0.500	0.772	0.500	0.688	0.500	0.768
TP	406	180	378	221	374	243	416	296	416	414	387	277	373	365	409	194	416	232	372	251
FP	155	8	120	24	99	30	164	23	6	0	86	19	67	63	162	13	167	32	103	23
FN	10	236	38	195	42	173	0	120	0	2	29	139	43	51	7	222	0	184	44	165
TN	12	159	47	143	68	137	3	144	161	167	81	148	100	104	5	154	0	135	64	144
Accuracy	0.717	0.582	0.729	0.624	0.758	0.652	0.719	0.755	0.990	0.997	0.803	0.729	0.811	0.805	0.710	0.597	0.714	0.630	0.748	0.678
Kappa	0.065	0.273	0.223	0.297	0.341	0.322	0.025	0.489	0.975	0.992	0.463	0.454	0.518	0.511	0.018	0.283	0.000	0.288	0.310	0.371
Sensitivity	0.976	0.433	0.909	0.531	0.899	0.584	1.000	0.712	1.000	0.995	0.930	0.666	0.897	0.877	0.983	0.466	1.000	0.558	0.894	0.603
Specificity	0.072	0.952	0.281	0.856	0.407	0.820	0.018	0.862	0.964	1.000	0.485	0.886	0.599	0.623	0.030	0.922	0.000	0.808	0.383	0.862
Precision	0.724	0.957	0.759	0.902	0.791	0.890	0.717	0.928	0.986	1.000	0.818	0.936	0.848	0.853	0.716	0.937	0.714	0.879	0.783	0.916
Negative Predictive Value	0.545	0.403	0.553	0.423	0.618	0.442	1.000	0.545	1.000	0.988	0.736	0.516	0.699	0.671	0.417	0.410	NA	0.423	0.593	0.466
Miss Rate	0.024	0.567	0.091	0.469	0.101	0.416	0.000	0.288	0.000	0.005	0.070	0.334	0.103	0.123	0.017	0.534	0.000	0.442	0.106	0.397
Fail-Out	0.928	0.048	0.719	0.144	0.593	0.180	0.982	0.138	0.036	0.000	0.515	0.114	0.401	0.377	0.970	0.078	1.000	0.192	0.617	0.138
False Discovery Rate	0.276	0.043	0.241	0.098	0.209	0.110	0.283	0.072	0.014	0.000	0.182	0.064	0.152	0.147	0.284	0.063	0.286	0.121	0.217	0.084
False Omission Rate	0.455	0.597	0.447	0.577	0.382	0.558	0.000	0.455	0.000	0.012	0.264	0.484	0.301	0.329	0.583	0.590	NA	0.577	0.407	0.534
Threat Score	0.711	0.425	0.705	0.502	0.726	0.545	0.717	0.674	0.986	0.995	0.771	0.637	0.772	0.762	0.708	0.452	0.714	0.518	0.717	0.572
F1 Score	0.831	0.596	0.827	0.669	0.841	0.705	0.835	0.805	0.993	0.998	0.871	0.778	0.871	0.865	0.829	0.623	0.833	0.682	0.835	0.728
Accuracy	0.717	0.581	0.729	0.624	0.758	0.652	0.719	0.755	0.990	0.997	0.803	0.729	0.811	0.804	0.710	0.597	0.714	0.630	0.748	0.678
Informedness	0.048	0.385	0.190	0.388	0.306	0.404	0.018	0.574	0.964	0.995	0.415	0.552	0.495	0.500	0.013	0.389	0.000	0.366	0.277	0.466
Balanced Accuracy	0.524	0.692	0.595	0.694	0.653	0.702	0.509	0.787	0.982	0.998	0.708	0.776	0.748	0.750	0.507	0.694	0.500	0.683	0.639	0.733

Table 6. Machine Learning Models Evaluation Metrics of USA Dataset

Models using 4-Fold Cross Validation	Logistic Regression	Logistic Regression	Naïve Bayes	Naïve Bayes	Naive Bayes	kNN	kNN	SVC	SVC	SVC	Random Forest (optimal ntree = 1550)	Random Forest (optimal ntree = 1550)	Gradient Boosting Machine	Gradient Boosting Machine	C 5.0	C 5.0	Feedforward Neural Network	Feedforward Neural Network	Model Averaged Neural Network	Model Averaged Neural Network	Multivariate Adaptive Regression Spline	Multivariate Adaptive Regression Spline
Cut-Off	0.5	0.464	0.5	0.435	0.448	0.5	0.448	0.5	0.422	0.5	0.888	0.888	0.5	0.359	0.5	0.406	0.5	0.481	0.5	0.385	0.5	0.516
TP	81	101	94	106	79	95	92	104	104	145	145	145	122	134	117	119	83	88	129	135	93	93
FP	40	58	28	37	35	53	30	43	43	0	0	0	16	31	14	16	27	31	8	11	28	24
FN	64	44	51	39	66	50	53	41	41	0	0	0	23	11	28	26	62	57	16	10	52	52
TN	160	142	172	163	147	165	170	157	157	200	200	200	184	169	186	184	173	169	192	189	172	176
Accuracy	0.6986	0.7043	0.771	0.7797	0.7014	0.7072	0.7594	0.7565	1	1	1	1	0.887	0.8783	0.8783	0.8783	0.742	0.7449	0.9304	0.9391	0.7681	0.7797
Kappa	0.367	0.401	0.519	0.547	0.381	0.389	0.495	0.301	1	1	1	1	0.767	0.755	0.747	0.748	0.452	0.463	0.856	0.875	0.513	0.536
Sensitivity	0.558621	0.696552	0.648276	0.731034	0.544828	0.635172	0.634483	0.717241	1	1	1	1	0.841379	0.924138	0.806897	0.82069	0.5272414	0.606897	0.889655	0.931034	0.641379	0.641379
Specificity	0.8	0.71	0.86	0.815	0.825	0.735	0.85	0.785	1	1	1	1	0.92	0.845	0.93	0.92	0.865	0.845	0.96	0.945	0.86	0.88
Precision	0.69421	0.63522	0.770492	0.741259	0.692982	0.641892	0.754098	0.707483	1	1	1	1	0.884058	0.812121	0.89313	0.881481	0.754545	0.739496	0.941606	0.924658	0.768595	0.794872
Negative Predictive Value	0.714286	0.763441	0.7713	0.806931	0.714286	0.714286	0.746193	0.762332	0.792929	1	1	1	0.888889	0.938889	0.869159	0.87619	0.73617	0.747788	0.923077	0.949749	0.767857	0.77193
Miss Rate	0.441379	0.303448	0.351724	0.268966	0.455172	0.344828	0.365517	0.282759	0	0	0	0	0.158621	0.075862	0.193103	0.17931	0.427586	0.393103	0.110345	0.068966	0.358621	0.358621
Fail-Out	0.2	0.29	0.14	0.185	0.175	0.265	0.15	0.215	0	0	0	0	0.08	0.155	0.07	0.08	0.135	0.155	0.04	0.053	0.14	0.12
False Discovery Rate	0.330579	0.36478	0.229508	0.258741	0.307018	0.338108	0.245902	0.292517	0	0	0	0	0.115942	0.187879	0.10687	0.118519	0.245455	0.260504	0.058394	0.075342	0.231405	0.205128
False Omission Rate	0.285714	0.236559	0.2287	0.193069	0.285714	0.253807	0.237668	0.207071	0	0	0	0	0.111111	0.061111	0.130841	0.12381	0.26383	0.252212	0.076923	0.050251	0.232143	0.22807
Threat Score	0.437838	0.497537	0.543353	0.582418	0.438889	0.479798	0.525714	0.535191	1	1	1	1	0.757764	0.761364	0.735849	0.73913	0.482558	0.5	0.843137	0.865385	0.537572	0.550296
F1 Score	0.609023	0.664474	0.70412	0.736111	0.610039	0.648464	0.689139	0.712329	1	1	1	1	0.862191	0.864516	0.847826	0.85	0.65098	0.666667	0.914894	0.927835	0.69248	0.709924
Informedness	0.358621	0.406552	0.508276	0.546034	0.369828	0.390172	0.484483	0.502241	1	1	1	1	0.761379	0.769138	0.736897	0.74069	0.437414	0.451897	0.849655	0.876034	0.501379	0.521379
Balanced Accuracy	0.67931	0.703276	0.754138	0.773017	0.684914	0.695086	0.742241	0.751121	1	1	1	1	0.88069	0.884569	0.868448	0.870345	0.718707	0.725948	0.924828	0.938017	0.75969	0.76069

Figure 6. ROC Curve for default and optimal threshold for Indian Dataset

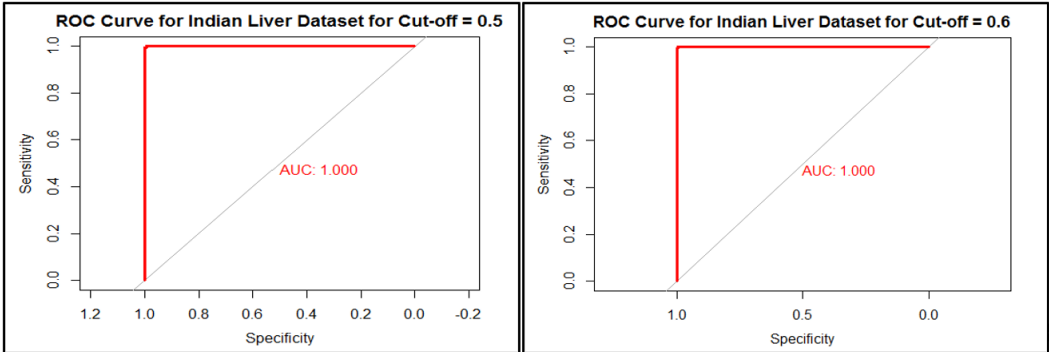
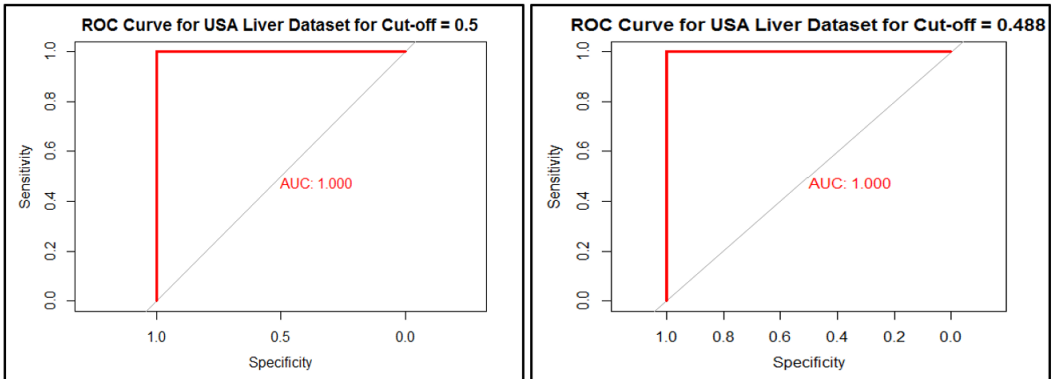


Figure 7. ROC Curve for default and optimal threshold for Indian Dataset



However, this study is limited by the sample sizes for both the datasets. The prediction accuracy can vary if the sample size increases. Apart from that, inclusion of other health indicators can also help the classification models to accurately identify liver disease in a patient. Studies based on different age groups as well as topologies can help to understand roles of different health indicators in liver disease detection. Future studies can focus on these areas to improve accuracy of the models.

5. CONCLUSION

Our findings from the analysis respond to the study's research questions and help to achieve its goals, which are to accurately diagnose the liver disease using classification algorithm and its parameters of measurement like sensitivity, accuracy, precision, kappa etc. The findings will help in automated classification algorithm selection and prediction of liver disease based on available features. This study will enable better decision making for medical practitioners and researchers in terms of liver disease detection. Predictive models will help patients in early diagnosis of liver related issues and disease. However, given the nature of datasets available, the study has to take care of overfitting issues due to use of advanced techniques and machine learning algorithms on clean and small size of samples. Figure 6 and 7 show the AUC (Area under the ROC curve) to be exactly 100% which is pertinent to these samples. The study proposes machine learning models

with extensive set of parameters and tuning using cross validation, Youden's index and algorithm specific parameters to achieve highest level of accuracy in liver disease prediction. Introduction of medically approved parameters in predictions would help in better understanding of the model in prediction. Medical practitioners will be able to make an informed decision with the help of intelligent detection system using predictive modelling. This will help to reduce health hazards and events like deaths significantly in liver disease diagnosis.

REFERENCES

- Alfisahrin, S. N. N., & Mantoro, T. (2013, December). Data mining techniques for optimization of liver disease classification. In *2013 International Conference on Advanced Computer Science Applications and Technologies* (pp. 379-384). IEEE. doi:10.1109/ACSAT.2013.81
- AnujaPriyama, A., Gupta, R., Ratheeb, A., & Srivastava, S. (n.d.). Comparative Analysis of Decision Tree Classification Algorithms. *International Journal of Current Engineering and Technology*.
- Arshad, I., Dutta, C., Choudhury, T., & Thakral, A. (2018, June). Liver Disease detection due to excessive alcoholism using Data Mining Techniques. In *2018 International Conference on Advances in Computing and Communication Engineering (ICACCE)* (pp. 163-168). IEEE. doi:10.1109/ICACCE.2018.8441721
- Baitharu, T. R., & Pani, S. K. (2016). Analysis of data mining techniques for healthcare decision support system using liver disorder dataset. *Procedia Computer Science*, 85, 862–870. doi:10.1016/j.procs.2016.05.276
- Buuren, S. V., & Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 1–68.
- Cohen, J. A., & Kaplan, M. M. (1979). The SGOT/SGPT ratio—An indicator of alcoholic liver disease. *Digestive Diseases and Sciences*, 24(11), 835–838. doi:10.1007/BF01324898 PMID:520102
- Durai, V., Ramesh, S., & Kalthireddy, D. (2019). Liver disease prediction using machine learning. *Int. J. Adv. Res. Ideas Innov. Technol.*, 5(2), 1584–1588.
- Faisal, M. I., Bashir, S., Khan, Z. S., & Khan, F. H. (2018, December). An evaluation of machine learning classifiers and ensembles for early stage prediction of lung cancer. In *2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST)* (pp. 1-4). IEEE. doi:10.1109/ICEEST.2018.8643311
- Farahnakian, F., Pahikkala, T., Liljeberg, P., & Plosila, J. (2013, December). Energy aware consolidation algorithm based on k-nearest neighbor regression for cloud data centers. In *2013 IEEE/ACM 6th International Conference on Utility and Cloud Computing* (pp. 256-259). IEEE. doi:10.1109/UCC.2013.51
- Fitriyah, H., & Setyawan, G. E. (2018, October). Automatic Estimation of Human Weight From Body Silhouette Using Multiple Linear Regression. In *2018 5th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)* (pp. 749-752). IEEE. doi:10.1109/EECSI.2018.8752763
- Friedman, J. H. (1991). Multivariate adaptive regression splines. *Annals of Statistics*, 1–67.
- Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics (Oxford, England)*, 16(10), 906–914. doi:10.1093/bioinformatics/16.10.906 PMID:11120680
- Geng, X. X., Huang, R. G., Lin, J. M., Jiang, N., & Yang, X. X. (2016). Transient elastography in clinical detection of liver cirrhosis: A systematic review and meta-analysis. *Saudi Journal of Gastroenterology: Official Journal of the Saudi Gastroenterology Association*, 22(4), 294. doi:10.4103/1319-3767.187603 PMID:27488324
- Gupta, S., Shrivastava, N. A., Khosravi, A., & Panigrahi, B. K. (2016, July). Wind ramp event prediction with parallelized gradient boosted regression trees. In *2016 International Joint Conference on Neural Networks (IJCNN)* (pp. 5296-5301). IEEE. doi:10.1109/IJCNN.2016.7727900
- Harper, P. R. (2005). A review and comparison of classification algorithms for medical decision making. *Health Policy (Amsterdam)*, 71(3), 315–331. doi:10.1016/j.healthpol.2004.05.002 PMID:15694499
- Huang, L. C., Hsu, S. Y., & Lin, E. (2009). A comparison of classification methods for predicting Chronic Fatigue Syndrome based on genetic data. *Journal of Translational Medicine*, 7(1), 81. doi:10.1186/1479-5876-7-81 PMID:19772600
- Jain, D., & Singh, V. (2020). A Novel Hybrid Approach for Chronic Disease Classification. *International Journal of Healthcare Information Systems and Informatics*, 15(1), 1–19. doi:10.4018/IJHISI.2020010101
- Jin, H., Kim, S., & Kim, J. (2014). Decision factors on effective liver patient data prediction. *International Journal of Bio-Science and Bio-Technology*, 6(4), 167-178.

- Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., & Klein, M. (2002). *Logistic regression*. Springer-Verlag.
- Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5), 1–26. doi:10.18637/jss.v028.i05
- Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software*, 36(11), 1–13. doi:10.18637/jss.v036.i11
- Kuzhippallil, M. A., Joseph, C., & Kannan, A. (2020, March). Comparative Analysis of Machine Learning Techniques for Indian Liver Disease Patients. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)* (pp. 778-782). IEEE. doi:10.1109/ICACCS48705.2020.9074368
- Li, B. N., Chui, C. K., Chang, S., & Ong, S. H. (2012). A new unified level set method for semi-automatic liver tumor segmentation on contrast-enhanced CT images. *Expert Systems with Applications*, 39(10), 9661–9668. doi:10.1016/j.eswa.2012.02.095
- Lin, R. H. (2009). An intelligent model for liver disease diagnosis. *Artificial Intelligence in Medicine*, 47(1), 53–62. doi:10.1016/j.artmed.2009.05.005 PMID:19540738
- Lin, R. H., & Chuang, C. L. (2010). A hybrid diagnosis model for determining the types of the liver disease. *Computers in Biology and Medicine*, 40(7), 665–670. doi:10.1016/j.combiomed.2010.06.002 PMID:20591425
- Mazaheri, P., Norouzi, A., & Karimi, A. (2015). Using algorithms to predict liver disease Classification. *Electronics Information and Planning*, 3, 255–259.
- Onak, O. N., Dogrusoz, Y. S., & Weber, G. W. (2017, September). Effect of the geometric inaccuracy in multivariate adaptive regression spline-based inverse ECG solution approach. In *2017 Computing in Cardiology (CinC)* (pp. 1-4). IEEE.
- Park, J., Kim, K. Y., & Kwon, O. (2014, August). Comparison of machine learning algorithms to predict psychological wellness indices for ubiquitous healthcare system design. In *Proceedings of the 2014 International Conference on Innovative Design and Manufacturing (ICIDM)* (pp. 263-269). IEEE. doi:10.1109/IDAM.2014.6912705
- Pathan, A., Mhaske, D., Jadhav, S., Bhondave, R., & Rajeswari, K. (2018). Comparative Study of Different Classification Algorithms on ILPD Dataset to Predict Liver Disorder. *International Journal for Research in Applied Science and Engineering Technology*, 6(2), 388–394. doi:10.22214/ijraset.2018.2056
- Priyanga, P., & Naveen, N. C. (2018). Analysis of Machine Learning Algorithms in Health Care to Predict Heart Disease. *International Journal of Healthcare Information Systems and Informatics*, 13(4), 82–97. doi:10.4018/IJHISI.2018100106
- Rajeswari, P., & Reena, G. S. (2010). Analysis of liver disorder using data mining algorithm. *Global Journal of Computer Science and Technology*.
- Ramana, B. V., Babu, M. S. P., & Venkateswarlu, N. B. (2011). A critical study of selected classification algorithms for liver disease diagnosis. *International Journal of Database Management Systems*, 3(2), 101–114. doi:10.5121/ijdm.2011.3207
- Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* (Vol. 3, No. 22, pp. 41-46). Academic Press.
- Schiff, E. R., Maddrey, W. C., & Reddy, K. R. (Eds.). (2017). *Schiff's diseases of the liver*. John Wiley & Sons. doi:10.1002/9781119251316
- Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study. *American Journal of Epidemiology*, 179(6), 764–774. doi:10.1093/aje/kwt312 PMID:24589914
- Singh, A., & Pandey, B. (2016). Diagnosis of Liver Disease by Using Least Squares Support Vector Machine Approach. *International Journal of Healthcare Information Systems and Informatics*, 11(2), 62–75. doi:10.4018/IJHISI.2016040104
- Singh, A., & Pandey, B. (2017). A KLD-LSSVM based computational method applied for feature ranking and classification of primary biliary cirrhosis stages. *International Journal of Computational Biology and Drug Design*, 10(1), 24–38. doi:10.1504/IJCBDD.2017.082807

- Sorich, M. J., Miners, J. O., McKinnon, R. A., Winkler, D. A., Burden, F. R., & Smith, P. A. (2003). Comparison of linear and nonlinear classification algorithms for the prediction of drug and chemical metabolism by human UDP-glucuronosyltransferase isoforms. *Journal of Chemical Information and Computer Sciences*, 43(6), 2019–2024. doi:10.1021/ci034108k PMID:14632453
- Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, 62(1), 77–89. doi:10.1016/S0034-4257(97)00083-7
- Sug, H. (2012). Improving the prediction accuracy of liver disorder disease with oversampling. *Applied Mathematics in Electrical and Computer Engineering*, 7, 331–335.
- Tietze, K. J. (2011). *Clinical skills for pharmacists-E-book: A patient-focused approach*. Elsevier Health Sciences.
- Vijayarani, S., & Dhayanand, S. (2015). Liver disease prediction using SVM and Naïve Bayes algorithms. *International Journal of Science, Engineering and Technology Research*, 4(4), 816–820.
- Wang, P., Jiang, T., Fan, G., & Dan, C. (2015, August). Prediction of Torpedo Initial Velocity Based on Random Forests Regression. In *2015 7th International Conference on Intelligent Human-Machine Systems and Cybernetics* (Vol. 1, pp. 337-339). IEEE. doi:10.1109/IHMSC.2015.17
- Wu, Q., & Zhao, W. (2017, October). Small-cell lung cancer detection using a supervised machine learning algorithm. In *2017 International Symposium on Computer Science and Intelligent Controls (ISCSIC)* (pp. 88-91). IEEE. doi:10.1109/ISCSIC.2017.22
- Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1), 32–35. doi:10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3 PMID:15405679
- Young, J. (2017). *Imputation for Random Forests*. Academic Press.

Aritra Pan is Assistant Professor and Chairperson at IMT Ghaziabad in the area of Business Analytics. Dr. Pan has done his Ph.D. from IIT Kharagpur in the area of Financial Markets Analytics. Dr. Pan was previously associated with global corporate firms like PricewaterhouseCoopers (PwC) and RS Software. Dr. Pan has extensively worked in the areas of Data Science, Data Strategy Consulting, Business Analytics and Business Intelligence.

Subrata Samanta is a Data Scientist & NLP engineer at PWC in the area of retail, banking, healthcare, tax & energy. Subrata has done his B.Tech from West Bengal University of Technology in Computer Science & Engineering. Subrata was previously associated with global analytics firms like Accenture and TCS. Subrata has expertise in the areas of Data Science, Machine Learning, Deep learning & NLP.