International Journal of Healthcare Information Systems and Informatics
Volume 17 • Issue 2

Liver Disease Detection:
Evaluation of Machine Learning Algorithms Performances With Optimal Thresholds

Aritra Pan, Indian Institute of Management, Bodh Gaya, India*
https://orcid.org/0000-0001-5740-0804

Shameek Mukhopadhyay, The Heritage Academy, Kolkata, India
https://orcid.org/0000-0002-5764-2368

Subrata Samanta, Ernst and Young, India

ABSTRACT

Intelligent predictive systems are showing a greater level of accuracy and effectiveness in early detection of critical diseases like cancer and liver and lung disease. Predictive models assist medical practitioners in identifying the diseases based on symptoms and health indicators like hormones, enzymes, age, bloodcounts, etc. This study proposes a framework to use classification models to accurately detect chronic liver disease by enhancing the prediction accuracy through cutting-edge analytics techniques. The article proposes an enhanced framework on the original study by Ramana et al. It uses evaluation measures like precision and balanced accuracy to choose the most efficient classification algorithm in India 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. A framework is proposed for highly accurate automated disease detection in the medical industry, and it helps in strategizing preventive measures for patients with liver diseases.

KEYWORDS

Balanced Accuracy. Classification Techniques, Liver Disease Detection, Precision, Youden Index

BACKGROUND

Liver is the largest glandular organ of the human body, which weighs around three pounds (Li et al., 2012). The liver performs different types of metabolic functions, like filtering blood, producing bile, assisting in fat digestion, making proteins for blood clotting, metabolising drugs, storing glucose and, most importantly, detoxifying harmful chemicals Singh et al. (2017). Malfunctioning of liver may cause liver disease and have serious health effects. The causes of liver disease are varied and can include consumption of contaminated food, inherited disorders, accumulation of excessive fat, hepatocytes damage due to infection by bacteria, viruses or fungi, and excessive consumption of alcohol or drugs Lin et al. (2010). Malnutrition, obesity leads to advanced stage of liver disease and

DOI: 10.4018/IJHISI.299956 *Corresponding Author

non-alcoholic fatty liver disease further leading to non-alcoholic steatohepatitis and cirrhosis for some cases as discussed by McClain et al. (2020). They have also discussed various causes and the methods for assessing malnutrition. A patient's survival rate may increase by several times if the diagnosis of the liver disease is done at an early stage, but diagnosis requires various examination tests by expert physicians. However, these do not always assure the correct diagnosis Takkar et. al (2017), but liver function tests do significantly help in examining liver disorders. The key parameters in these tests include albumin, alkaline phosphatase, total proteins, alanine aminotransferase, aspartate aminotransferase, direct bilirubin, total bilirubin, gamma-glutamyl transferase, prothrombin time, triglycerides and platelet counts. Liver diseases are categorised into more than 100 types, and the disease can be acute or chronic. Some liver disease has successful treatments, while others do not.

Liver disease and prediction now increasingly depend on intelligent systems, which now play a significant role in the medical industry. Data mining algorithms, neural networks and statistical techniques are widely applied to liver examination data to evaluate illnesses. Predictive modelling is a broadly used intelligent technique for automated detection of multiple diseases. Machine learning calculations provide specialists with essential measurements, continuous information and progressive examination data about a patient's illness, lab test results, preliminary clinical information, and family history. As identified by Jesty (2019) machine learning tasks in the field of medicine can be classified into categories namely (i) genomics which is the study of DNA, (ii) audio analysis that is interpretable pattern of digital audio recording, (iii) computer vision which extracts information from digital images and videos by running algorithms, (iv) natural language processing which process text for meaningful information and (v) health record regression that establishes relations between features. The guarantee for improving the detection and prediction of disease has increased interest in machine learning in the biomedical field and had improved the decision-making process by increasing its objectivity. Not only for liver but also for any disease the diagnosis involves many uncertainties in the information system and to handle them different types of intelligent techniques are used with a proposed model as shown used by P. and Acharjya (2020).

Classification algorithms are cost effective and can be implemented in different automated medical diagnosis tools. The aim of this study is to enhance the predictive models used in the original study by Ramana et al. (2011) to increase the accuracy and effectiveness of current models. The study focuses on two patient data sets (INDIA and USA), and we have applied various machine learning techniques like Logistic Regression, Naïve Bayes, kNN, SVC, Random Forest, Gradient Boosting Machine, C 5.0, Feedforward Neural Network, Model Averaged Neural Network, and Multivariate Adaptive Regression Spline. We have estimated the performance measures of these techniques from various perspectives, such as Balanced Accuracy, Precision, recall and F1 – Score, to propose an efficient classification algorithm for liver disease detection based on the levels of various enzymes, patient age and other factors. The best classification algorithm in terms of accuracy, precision, sensitivity, and specificity was identified by comparing the performance of each algorithm using the receiver operating characteristic (ROC) curve. This study aims to propose a framework to help medical practitioners and other concerned researchers to accurately diagnose liver using a broad set of evaluation parameters. The study findings will help medical practitioners to make better and accurate decisions in liver disease detection.

LITERATURE REVIEW

Accurate predictions of liver disease have been obtained implementing machine learning techniques by different researchers as a result of improvements in technology. The work done in analysing liver disease by machine learning are by way of analysis of images of MRI, CT or ultrasound scans, predicting mortality among hepatitis patients and by way of analysing numeric and binary data for diagnosis of liver disease. Machine learning can also be integrated with smart IoT devices for collecting patient's data accurately and those IoT devices need to have proper security protection

where machine learning can be a solution, Mohanta (2021). Also, the IoT devices can be used to gather physiological signals of the patients for further processing, which is supported by machine learning techniques, Banerjee et al. (2020). The machine learning techniques can be solved by using Nature Inspired algorithms as discussed by Kauser et al. (2017). Arshad et al. (2018) used data mining techniques to detect liver disease caused by excessive alcoholism. They constructed a decision tree for the datasets and used it to generate the rules. The source of the data was University of California at Irvine (UCI) laboratory and based on it the training dataset was developed. Statistical techniques, data mining algorithms and neural networks have also been widely deployed on liver examination data to evaluating liver disease. For instance, a data classification technique was proposed by Rajeswari and Reena (2010) where their experimental result dealt with data classification obtained from FT Tree algorithm, KStar algorithm and Naive Bayes algorithm. Their experimental result showed Function Trees provide 97.10% of correct result. Similarly, a proposal was made for identification of liver disease based on 10 important patient attributes using a decision tree, Naïve Bayes, and NB Tree algorithms by Alfisahrin and Montaro (2013), who designed a model in the Weka tool. Their model gave maximum accuracy with NB Tree algorithm whereas the minimum computation time was given by Naïve Bayes algorithm.

The concept of using various classification techniques to assist doctors in determining disease quickly and efficiently was described by Jinet al. (2014), who compared and analysed various classifiers, such as Naïve Bayes, Decision Tree, Multi-Layer Perceptron and k-NN, based on several parameters, including specificity and sensitivity. The algorithms were again implemented with the Weka, and the UCI Repository was used for the collection of the dataset. The outcome of the experiment was better classification results in terms of precision from Naïve Bayes and better recall and sensitivity from Logistic Regression and Random Forest.A study by Ramana et al. (2011) on various classification algorithms based on different attributes was carried out on different types of liver datasets and the performance was evaluated. SVM, Back propagation and KNN performed better on the selected dataset. Mazaheri et al. (2015) applied different classification algorithms in liver tumour classifications or segmentations. Durai et al. (2019) predicted liver diseases using machine learning and found that some of the machine learning approaches were not viable for a large volume of data. They determined that the classification process did not require large volumes of data and that the cohesion that a classifier shares with a particular set of data should stand did not need to be viable for the rest of the training set. They identified that the data quantity, features, and quality presented a major challenge for the accuracy of machine learning. For Indian liver disease patients, a comparative analysis of machine learning techniques was conducted by Kuzhippallil et al. (2020) and they proposed a method for building a predictive model for liver disease using various supervised machine learning algorithms. To get the best attributes for prediction of the liver disease they applied a genetic algorithm in combination with XGBoost, thereby effectively utilising various performance metrics. Recent studies from machine learning shows prediction of road accidents and thereby developing a smart transportation system using different machine learning models have been used by Mohanta et al. (2021) where the highest mean accuracy was achieved from Decision Tree. Artificial bee colony and rough set hybridization technique was used by Acharjya et. al (2021) which was applied on a hepatitis dataset and the proposed model helps in detection of the disease accurately with an accuracy of 96.2%.

For detection of malignant liver and its treatment, computer aided diagnosis method was adopted by Khan et al. (2022). Baitharu and Pani (2016) presented a method for the medical diagnoses of liver by learning pattern through the collected data. They used six popular classifier algorithms namely J48, Naive Bayes, ANN, ZeroR, 1BK and VFI and found improved predictive performance of all classifiers except Naive Bayes. A comparative study was performed by Pathan et al. [14] where 100% accuracy was found with Random Forest algorithm. The parameters for the study were accuracy, error rate, precision, recall and F-measure. Vijayarani and Dhayanand (2015) demonstrated a predictive analysis of liver disorder using two classification algorithms and the comparison was done on execution time

and accuracy. The result showed minimum execution time for Naïve Bayes classifier and highest classification accuracy for SVM classifier. Singh and Pandey (2016) presented an approach for diagnosis of liver disease by deploying various classification methods. Experimental results showed that Support Vector Machine based approaches have the best diagnostic accuracy rates and the best predictive model was least squares support vector machine. Sug (2012) performed an experiment on the BUPA liver dataset. By way of duplication, he increased the instances of minor classes so that the disdaining property is compensated while generating algorithms for decision tree as the dataset is relatively small and is having high error rate. Over sampling technique was used by him for minor classes and the experiment result showed good results with C4.5 and CART which are decision tree algorithms. He recommended them for oversampling for the data set to generate decision trees. For diagnosis of liver fibrosis and cirrhosis, Geng et al. (2016) used transient elastography where specificity was found to be 88%, sensitivity to be of 81%. For optimization and deriving optimal solutions in healthcare industry the hybrid nature-inspired algorithms have been used by Kauser et. Al (2020).

Overall, various studies have aimed at developing better prediction models for liver disorder disease. The use of data mining algorithms by various researchers is providing a better solution to the issue.

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 and artificial intelligence, in their efforts to improve diagnosis and treatment. This has created a scope for providing excellent medical solutions for patients. Healthcare management is one area which is broadly using predictive analytics for different objectives, like disease detection, patient care, patient recovery and drug formulation, as discussed by Park et al. (2014).

Liver disease, despite being one of the commonest diseases in the world, remains difficult to detect at early stages (Lin et al., 2009; Faisal et al., 2018; Wu et al., 2017). It can be triggered by many factors, like smoking, consumption of alcohol in excess, ingestion of arsenic-contaminated drinking water, obesity, low immunity, and inheritance, and it can be identified by analysis of several different enzymes in blood (Schiff et al., 2017). An efficient classification algorithm could help detection of liver disease early given the necessary patient data. Several studies (Sorich et al., 2003; Lin et al., 2009; Harper et al., 2005; Huang et al., 2009; Ramana et al., 2011; Wu et al., 2017; Faisal et al., 2018) have used in different types of machine learning algorithms, like Random Forest, Support Vector Machine (SVM), Neural Network, K nearest neighbour, Naïve Bayes, and Decision Tree, for the prediction of liver disease from chemical and medical datasets.

The present research examines the prediction of the presence of liver disease based on two datasets: an Indian database (ILPD; Indian Liver Patient Dataset) and the American UCI repository (Liver Disorders Data Set). This work is an enhancement of the groundwork of Ramana et al. (2011) and focuses on optimising the model and enhancing the liver disease detection using a broad set of parameters. In this research, we proposed to enhance the power of the predictive models for detecting liver disease in patients. Our goal was to improve the machine learning model of the original work done by Ramana et al. (2011) using different health indicators, like age and enzyme details. The findings will help the medical industry to improvise and to reduce errors in the identification of diseases and to arrive at accurate proposals for cures.

The literature reviewed above has provided an in-depth review of the techniques involved in critical health care detection of diseases like liver disease. The present article has framed following objectives for examination using a 4-fold cross validation on two datasets (Indian and US patients):

1. Predict and compare the accuracies of multiple machine learning models based on the original study by Ramana et al. (2011).
2. Use Youden's J Statistics to improve the classification models to yield the best prediction results.

3. Use feature engineering to remove the high correlation among factors affecting liver disease detection accuracy.
4. Optimise feature selection for the best performing model.
5. Use a grid search for model optimisation.

The findings will provide one of the primary analysis methods for understanding and predicting liver disease using a robust machine learning model.

METHODS

This study focuses on identifying a patient with liver disease based on selected attributes. We are proposing a classification model that is sufficiently capable to identify liver disease without any intervention from an expert doctor. This kind of model can be leveraged in the medical industry, where it will help medical practitioners to arrive at accurate diagnoses in a short time.

Data Description

In this study, liver disease detection has been optimised using different classification algorithms. The two liver patient datasets used in this study are samples from the Indian ILPD and the American Liver Disorders datasets collected from the UCI Machine Learning Repository. The attributes of Indian (INDIA) data set are age, gender, total bilirubin, direct bilirubin, alkaline phosphatase (alkphos), alanine aminotransferase (SGPT), aspartate aminotransferase (SGOT), total proteins, albumin and albumin and globulin Ratio (A/G ratio); these are shown in Table 1. The attributes of the American (USA) data set are mean corpuscular volume (mcv), alkaline phosphatase (alkphos), alanine aminotransferase (SGPT), aspartate aminotransferase (SGOT), gamma-glutamyl transpeptidase (gammagt), and number of half-pint equivalents of alcoholic beverages consumed per day (drinks) and are provided in Table 2. The liver functional tests common to both data sets are alkphos, SGPT and SGOT. The USA data set contains 345 patient records, and the INDIA data set contains 583 patient records. Figure 1 demonstrates the label-wise count distributions for the ratios of patients with disease vs without disease of 416:167 and 145:200 for the INDIA and USA datasets, respectively. Figure 1 shows the relative frequencies of dependent variables for the INDIA and USA datasets.

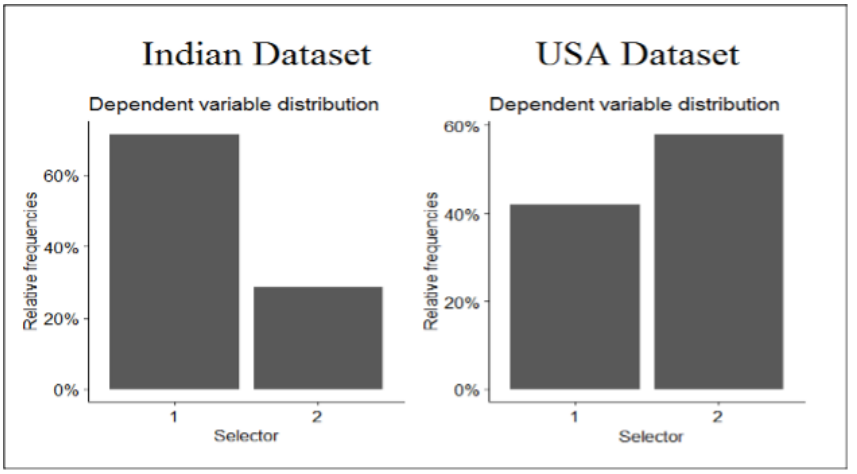
Table 1. Data description of Indian liver patient's dataset

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 phosphotase (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 phosphotase (alkphos)	69.87	18.34767
alanine aminotransferase (SGPT)	30.41	19.51231
aspartate aminotransferase(SGOT)	24.64	10.06449
gamma-glutamyltranspeptidase(gamma gt)	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 and 2 represents patients without liver disease)

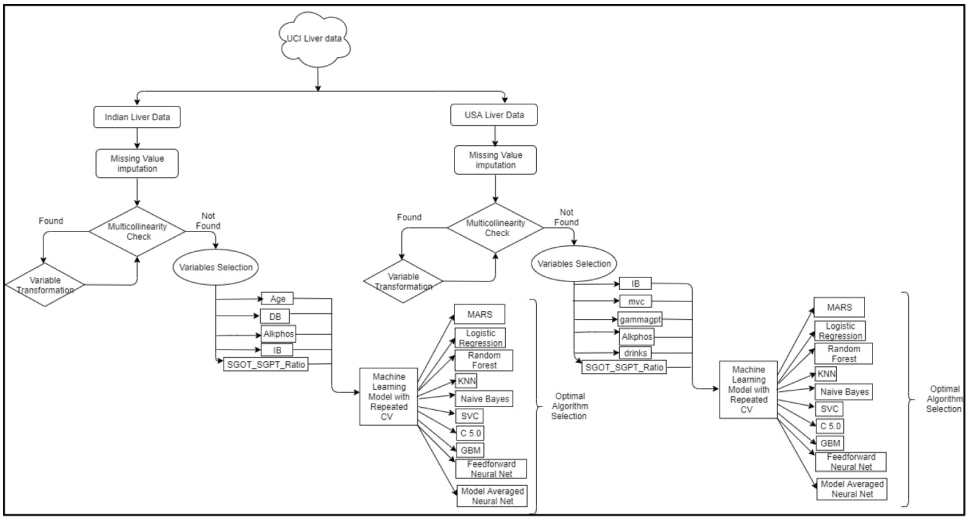


Proposed Framework for Study

We accomplished the task of the classification problem by following several steps, including data pre-processing, training machine learning model and tuning, to reach a conclusion. The INDIA/USA liver patient data were first collected from the UCI/UCLA repository and subjected to exploratory data analysis to investigate the presence of missing values and multi collinearity in the data. The missing values in the selected column (A.G. Ratio) were imputed to obtain the completeness of the data. Multicollinearity check was done in data before modelling. In the Indian data, variable transformations were done to counter its effect. A similar process was also followed for missing value handling, multi collinearity checks and feature engineering for the USA data before training the machine learning models. A variable selection process was then performed based on the relative importance of the variables in presence of dependent variable to make our model easier to understand and interpret. The data were then considered ready for application to the machine learning algorithms.

We performed repeated cross validations while training the different machine learning algorithms on the training data set. We then analysed the F1 scores of all the machine learning algorithms to select the optimal classifier. Figure 2 shows a flowchart of the entire process.

Figure 2. Process Flow



Missing Value Handling

The INDIA dataset was complicated by the issue of missing values, as indicated in Table 3. We used a random forest algorithm to fill in the missing values (Young, 2017). The Mice package (Buuren et al., 2011) in R is 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). To deal with continuous missing values (variables), Mice with random forests assigns the missing values by applying random produces from independent normal distributions which are centred on the means predicted from random forests. The out-of-the-bag mean square of the error is used as an estimator of the residual variance (Young, 2017). USA dataset does not have any missing values as shown in the Table 4.

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%
gamma gt	0%
drinks	0%

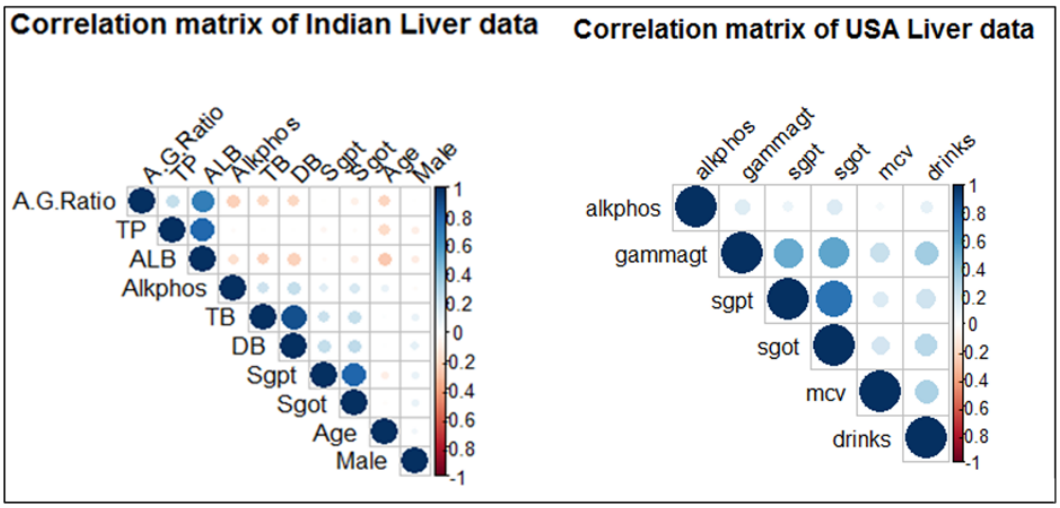
Feature Engineering and Correlation Analysis

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 the INDIA dataset. To mitigate the issue of high correlation, we created two new variables. In the USA dataset, we introduced a ratio between SGOT and SGPT, which is a significant parameter of identifying liver disease [30]. For the other two correlated variables direct bilirubin (DB) and total bilirubin (TB), we have computed indirect bilirubin (IB) by subtracting DB from TB, which is a common practice by pharmacists (Tietze, 2011). These feature additions removed the correlation from the data after removal of TB, SGOT and SGPT from their respective datasets.

Feature Selection

To extract useful insights from high volume data, statistical techniques are needed to reduce the noise or redundant data. Since, it is not necessary for every feature to be used to train a model. Models can be improved by only including uncorrelated and non-redundant attributes. That's why feature selection is necessary. Not only does it help in faster training of the model, but it also reduces the complexity of the model, makes the model and results interpretation comparatively easy as well as overall improves the performance metrics.

Figure 3. Correlation analysis of Indian and USA dataset



We have used Boruta algorithm (Kursa et al., 2010), where the shadow features get created. The Boruta algorithm works as a wrapper built around the random forest classification algorithm. It tries to capture all the important features in the dataset with respect to a dependent (outcome) variable. Using the Boruta algorithm, we duplicated the dataset and shuffled the values in each column. We then ran random forest classifiers on the merged dataset to estimate the variable importance of each feature based on mean decrease accuracy measure. The maximum Z score (MZSA) among the shadow was calculated and we tagged the variables as 'important' (Figure 4 and 5) if they had notably higher

Figure 4. Feature selection of Indian liver patient dataset

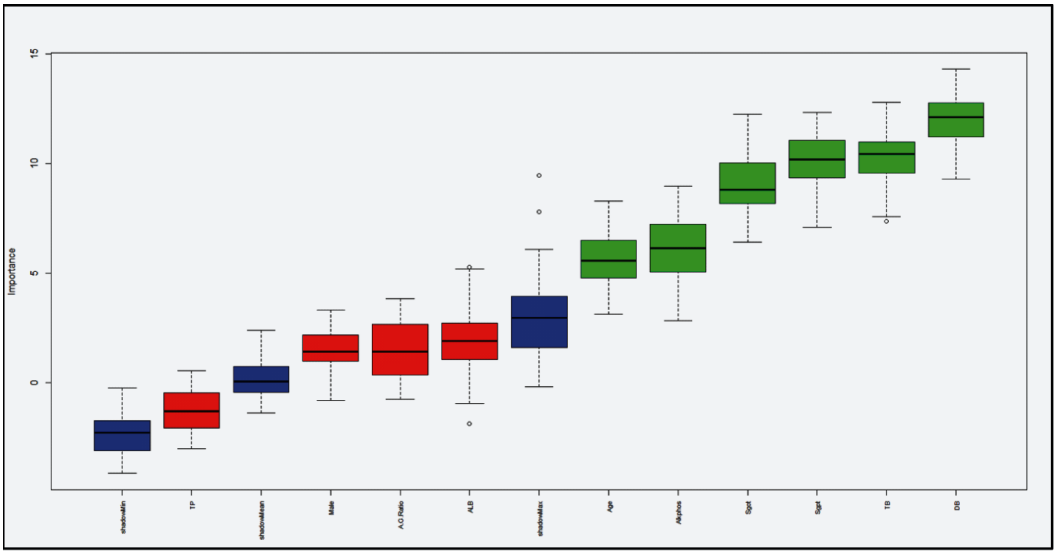
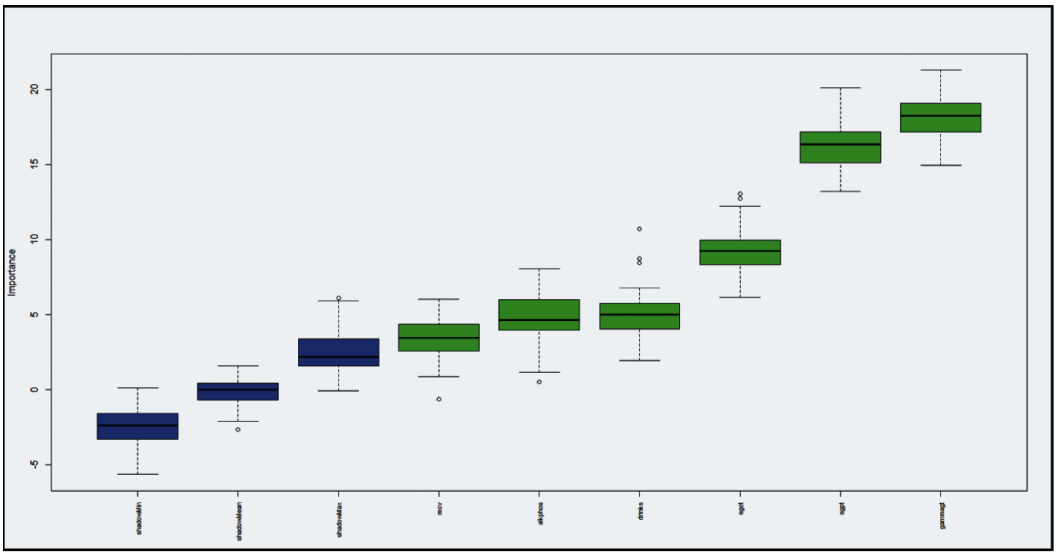


Figure 5. Feature selection of USA liver patient dataset



importance than the MZSA. We repeated the same steps for predefined number of iterations until all features were tagged with one category.

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

Machine Learning Algorithms

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

Indian dataset:

$$\text{Selector} \sim \text{Age} + \text{DB} + \text{Alkphos} + \text{IB} + \text{SGOT\_SGPT\_Ratio} \tag{1a}$$

USA dataset:

$$\text{Selector} \sim \text{mcv} + \text{Alkphos} + \text{gammagt} + \text{drinks} + \text{SGOT\_SGPT\_Ratio} \tag{1b}$$

This study has considered following machine learning algorithms for modelling: Logistic Regression, Naïve Bayes, kNN, SVC, Random Forest, Gradient Boosting Machine, C 5.0, Feedforward Neural Network, Model Averaged Neural Network, and Multivariate Adaptive Regression Spline.

Performance Evaluation of Machine Learning Algorithm

This research uses classification models like Logistic Regression, Naïve Bayes, kNN, SVC, Random Forest, Gradient Boosting Machine, C 5.0, Feedforward Neural Network using BFGS optimization, Model Averaged Neural Network, and Multivariate Adaptive Regression Spline using 4 – fold cross validation ($k = 4$) and grid search parameters. Grid searches for a 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 random forest, we have detected optimal number of trees for decision making in each case. 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 the confusion matrix (Stehman, 1997). We have computed different parameters, namely 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 patient points of each class individually.
- **Informedness:** Informedness measures how our model is informed about positive and negative 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 forecasted 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 positives. In our case, the false discovery rate reflects the situation when the model predicted someone as a liver patient, but the patient does not actually have the disease.
- **False Omission Rate:** The false omission rate measures the proportion of false negatives which are incorrectly rejected. In our case, this indicates the number of wrongly rejected predictions termed as non-liver patients but when the patient is a liver patient.

After achieving the values using the default classification threshold of 0.5, we tuned the model using Youden's J Statistics or Youden's Index [45] to achieve maximum Sensitivity and Specificity to maximise detection accuracy. This method helped the models to gain maximum Informedness in terms of probability of an informed decision.

RESULTS

The primary objective of this study was to accurately identify the presence of liver disease in a patient. The liver disease datasets INDIA and USA were taken from the UCI machine learning repository. These datasets reflect the presence or absence of liver disorder in patients based on the various medical test results per formed on the patients. The key features from the tests for the INDIA patients included age, total bilirubin, direct bilirubin, albumin and globulin ratio, alkaline phosphatase, albumin, alanine aminotransferase, aspartate aminotransferase and total proteins. For the USA dataset, the key features were the mean corpuscular volume, alkaline phosphatase, alaine aminotransferase, aspartate aminotransferase, gamma-glutamyl transeptidase and alcoholic beverages drunk per day. The INDIA dataset contains 583 samples belonging to two distinct classes (416 or 71.35% are cases of patients with liver disorder and 167 or 28.65% are healthy individuals). For USA, the sample size is 345, with 42% being patients with liver disorder.

The classification methods implemented in the study were Logistic Regression, Random Forest, K-Nearest Neighbour (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). These approaches included a four – fold cross validation technique and optimisation of the models with Youden's index. The experimental results in Table 5 and Table 6 confirmed that all the algorithms showed significant prediction performance improvements after tuning with Youden's index. The study considers Balanced Accuracy as a better evaluation metric compared to Accuracy. Nevertheless, random forest, which was tuned for the optimal number of trees based on minimal OBB error, showed the best accuracy rates with the optimised cut-off value using Youden's Index for both the INDIA and USA liver datasets. We have taken a range of 5 to 2000 trees to grow the random forest to select the best random forest model. The Model Averaged Neural Network showed significant improvement for the results in the USA dataset, which was designed using 5input, 1 hidden and one output layer in its structure.

This study focused on a broad set of evaluation metrics for identification of the optimal classification algorithm, in contrast to previous studies made on liver disease identification (Ramana et al., 2011; Takkar et al., 2017), chronic disease classification (Jain et al., 2020) and heart disease prediction (Priyanga et al., 2018). Using Youden's Index, we were able to increase the precision of the models significantly compared to values against a default threshold of 0.5. The use of Youden's Index helped the models to increase the probability of an informed decision, thereby reducing the negative predictive value and false positive rate (Table 5 and 6). The numbers of false positives were significantly reduced and increased the precision of the models. In the INDIA data, almost all the model's accuracy decreased, except for the random forest and support vector classification, where the accuracy increased after use of the cut-off value from the Youden's Index.

DISCUSSION

This study introduced data engineering techniques which were applied to both the INDIA and USA liver datasets before applying machine learning algorithms. The missing values found in the INDIA dataset were imputed based on predicted values using the random forest technique. This eliminated the 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, we formulated new explanatory variables by conducting feature engineering on highly correlated independent variables.

Table 5. Machine Learning Models Evaluation Metrics of Indian Dataset

Models using 4-Fold Cross Validation	Logistic Regression	Logistic Regression	Naïve Bayes	Naïve 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
Fall-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	kNN	kNN	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	0.5	0.448	0.5	0.422	0.5	0.488	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	145	145	122	134	117	119	83	88	129	135	93	93
FP	40	58	28	37	35	53	30	43	0	0	16	31	14	16	27	31	8	11	28	24
FN	64	44	51	39	66	50	53	41	0	0	23	11	28	26	62	57	16	10	52	52
TN	160	142	172	163	165	147	170	157	200	200	184	169	186	184	173	169	192	189	172	176
Accuracy	0.6986	0.7043	0.771	0.7797	0.7072	0.7014	0.7594	0.7565	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	0.501	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	0.655172	0.634483	0.717241	1	1	0.841379	0.924138	0.806897	0.82069	0.572414	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	0.92	0.845	0.93	0.92	0.865	0.845	0.96	0.945	0.86	0.88
Precision	0.669421	0.63522	0.770492	0.741259	0.692982	0.641892	0.754098	0.707483	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	0.746193	0.762332	0.792929	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	0.158621	0.075862	0.193103	0.17931	0.427586	0.393103	0.110345	0.068966	0.358621	0.358621
Fall-Out	0.2	0.29	0.14	0.185	0.175	0.265	0.15	0.215	0	0	0.08	0.155	0.07	0.08	0.135	0.155	0.04	0.055	0.14	0.12
False Discovery Rate	0.330579	0.36478	0.229508	0.258741	0.307018	0.358108	0.245902	0.292517	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	0.111111	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	0.553191	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	0.862191	0.864516	0.847826	0.85	0.65098	0.666667	0.914894	0.927835	0.699248	0.709924
Informedness	0.358621	0.406552	0.508276	0.546034	0.369828	0.390172	0.484483	0.502241	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	0.88069	0.884569	0.868448	0.870345	0.718707	0.725948	0.924828	0.938017	0.75069	0.76069

The formulation of direct DB, indirect DB and the SGOT – SGPT ratio were on par with the medical literature values (Cohen et al., 1979; Tietze, 2011). This is one of the unique approaches followed by this study that was not present in the base study by Ramana et al. (2011), as well as in the other literature (Takkar et al., 2017; Singh et al., 2016).

In results, the Random Forest algorithm surpassed all other algorithms in terms of performance for both the datasets. Informedness saw a significant increase in SVC compared to all other models in the INDIA liver dataset. The approach using Youden's index resulted in significant improvement of precision which also ensured effectiveness in identifying a patient with liver disease. However, this was found for the INDIA dataset only. For the USA dataset, we saw an improvement in the negative predictive value which helps in identifying healthy patients by reducing number of false negatives in identification. Overall, however, the balanced accuracy improved significantly for both datasets after the use of Youden's index. Apart from the random forest, the C 5.0 and GBM showed significant accuracy in predictions. Informedness also increased in all the models with Youden's index compared to the default threshold value 0.5.

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

CONCLUSION

Our findings from the analysis answer the study's research questions and help to achieve its goals, which are to accurately diagnose liver disease using a classification algorithm and to determine the algorithm parameters of measurement like sensitivity, accuracy, precision, kappa etc. The study findings will enable better decision making by 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 the datasets available, care must be taken to avoid overfitting issues due to the use of advanced techniques and machine learning algorithms on clean and small-sized samples. Figures 6 and 7 show the AUC (area under the ROC curve) to be exactly 100%, which is pertinent to these samples. We propose machine learning models with an extensive set of parameters and tuning using cross validation, Youden's index and algorithm-specific parameters to achieve the highest level of accuracy in liver disease prediction. Introduction of medically approved parameters in the predictions would help in better

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

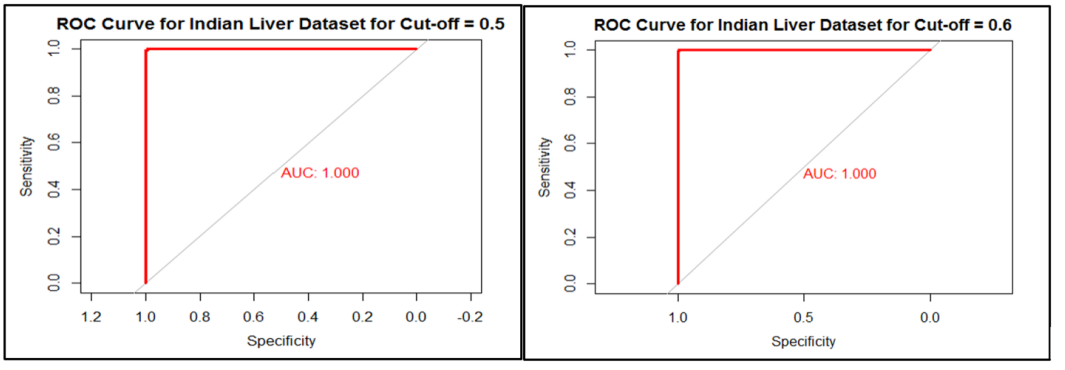
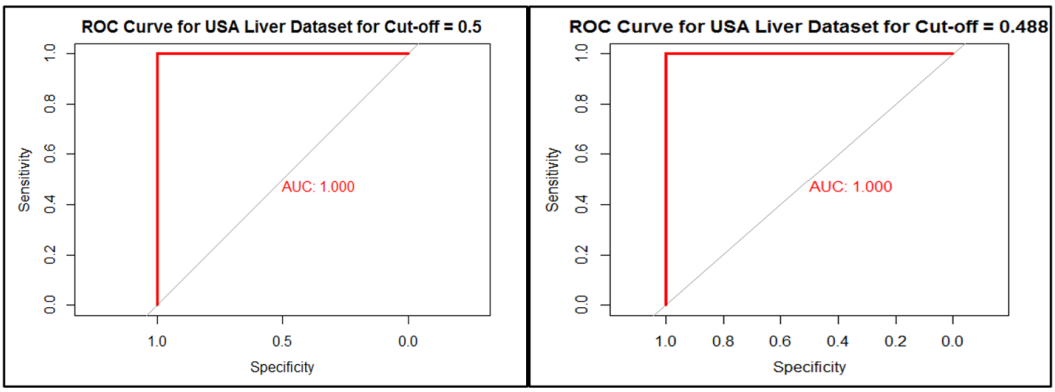


Figure 7. ROC Curve for default and optimal threshold for USA Dataset



understanding the model. Medical practitioners will be able to make an informed decision with the help of an intelligent detection system using predictive modelling. This will help to significantly reduce health hazards and events like deaths in liver disease diagnosis.

FUNDING AGENCY

Publisher has waived the Open Access publishing fee.

REFERENCES

Acharjya, D. P. (2021). Knowledge Inferencing Using Artificial Bee Colony and Rough Set for Diagnosis of Hepatitis Disease. *International Journal of Healthcare Information Systems and Informatics*, *16*(2), 49–72. doi:10.4018/IJHISI.20210401.oa3

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

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

Banerjee, A., Mohanta, B. K., Panda, S. S., Jena, D., & Sobhanayak, S. (2020, January). A secure IoT-fog enabled smart decision making system using machine learning for intensive care unit. In *2020 International Conference on Artificial Intelligence and Signal Processing (AISP)* (pp. 1-6). IEEE. doi:10.1109/AISP48273.2020.9073062

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, 22*(4), 294.

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), 1–8. 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

Jesty, B. (2019). *Machine learning for liver disease classification* (Doctoral dissertation). University of Southampton.

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.

Kauser, A. P., & Agrawal, R. (2020). Cluster Analysis of Health Care Data Using Hybrid Nature-Inspired Algorithms. *Recent Advances in Hybrid Metaheuristics for Data Clustering*, 101-111.

Khan, R. A., Luo, Y., & Wu, F. X. (2021). Machine learning based liver disease diagnosis: A systematic review. *Neurocomputing*. Advance online publication. doi:10.1016/j.neucom.2021.08.138

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*(1), 1–26. PMID:19777145

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.compbiomed.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.

McClain, C. J., Smart, L., Safadi, S., & Kirpich, I. (2020). Liver disease. In *Present Knowledge in Nutrition* (pp. 483–502). Academic Press. doi:10.1016/B978-0-12-818460-8.00026-5

Mohanta, B. K., Jena, D., Mohapatra, N., Ramasubbareddy, S., & Rawal, B. S. (2021). Machine learning based accident prediction in secure iot enable transportation system. *Journal of Intelligent & Fuzzy Systems*, 1-13.

Mohanta, B. K., Satapathy, U., & Jena, D. (2021). Addressing Security and Computation Challenges in IoT Using Machine Learning. In *Advances in Distributed Computing and Machine Learning* (pp. 67–74). Springer. doi:10.1007/978-981-15-4218-3_7

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.

P, K. A., & N, S. K. (2018). A Comprehensive Review of Nature-Inspired Algorithms for Feature Selection. In S. Dash, B. Tripathy, & A. Rahman (Eds.), *Handbook of Research on Modeling, Analysis, and Application of Nature-Inspired Metaheuristic Algorithms* (pp. 331-345). IGI Global. 10.4018/978-1-5225-2857-9.ch016

P., K.A., & Acharjya, D. P. (2020). A hybrid scheme for heart disease diagnosis using rough set and cuckoo search technique. *Journal of Medical Systems*, *44*(1), 1-16.

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

Priyam, A., Abhijeeta, G. R., Rathee, A., & Srivastava, S. (2013). Comparative analysis of decision tree classification algorithms. *International Journal of Current Engineering and Technology, 3*(2), 334-337.

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/ijdms.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.

Takkar, S., Singh, A., & Pandey, B. (2017). Application of machine learning algorithms to a well defined clinical problem: Liver disease. *International Journal of E-Health and Medical Communications*, *8*(4), 38–60. doi:10.4018/IJEHMC.2017100103

Tietze, K. J. (2011). *Clinical skills for pharmacists-E-book: A patient-focused approach*. Elsevier Health Sciences.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(1), 1–67. doi:10.18637/jss.v045.i03 PMID:27818617

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 (AP) is an Assistant Professor at Indian Institute of Management Bodh Gaya in the area of IT Systems and Analytics. Aritra has completed his Ph.D. from IIT Kharagpur, and, M.Tech. and B.E. from IIEST, Shibpur. Prior to joining Indian Institute of Management Bodh Gaya, Aritra was associated with IFMR Graduate School of Business (Krea University), IMT Ghaziabad and PwC. Post his Ph.D., Aritra worked as a Lead Data Scientist in PwC where he managed a team of Data Science and Data Analytics Consultants in Management Consulting (Advisory) vertical. Aritra has extensively worked in the areas of Data Science, Business Analytics and Business Intelligence during his association with global firms like PwC, RS Software etc.

Shameek Mukhopadhyay (SM) has M.Tech, MBA, B.Tech with 14 years of work experience in academics and industry. Shameek is interested in Data analytics, marketing.

Subrata Samanta (SS) is associated with Ernst & Young Kolkata. He is working as Data Scientist and NLP engineer in the area of retail, banking, healthcare, tax and energy. He has prior expertise in the areas of Data Science, Machine Learning, Deep learning and NLP.