Liver Disease Prediction using Evolutionary Hypertuned Ensemble Algorithm

*Milan Agrawal[1], Muskaan Chhikara[2], Yashika[3]

[1]CSE Department, IGDTUW, Delhi, India
[1] milan010btcse19@igdtuw.ac.in

[2] CSE Department, IGDTUW, Delhi, India
[2] muskaan064btcse19@igdtuw.ac.in

[3] CSE Department, IGDTUW, Delhi, India
[3] yashika069btcse19@igdtuw.ac.in

*Dr. Kalpana Yadav[1], Dr. Rishabh Kaushal[2]

[1] CSE Department, IGDTUW, Delhi, India

[2] CSE Department, IGDTUW, Delhi, India

Abstract: Due to a sedentary way of life, lifestyle diseases have become more frequent. Traditional methods involving physical examination and lab results make it difficult to diagnose. Thus, the application of machine learning models comes to aid by facilitating the early diagnosis of liver diseases. We focus on liver disease by exploring various machine learning (ML) algorithms, namely, Random Forest(RF), Support Vector Machine (SVM), Multi-Layer Perceptron (MLP) neural network, and Bayesian Classifier. We use hyperparameter tuning with the machine learning models to find an optimal combination of hyperparameters. Genetic Algorithm (GA) approach has been used to optimize Random Forest, SVM and Bayesian Classifier. We obtain average accuracies as 70%, 70%, 72%, 61%, 87.42%, 73%, 71% for RF, SVM, MLP Neural Networks, Bayesian Classifier, GA-Random Forest (GA-RF), GA-SVM and GA-Bayesian Classifier, respectively. Among these models Random Forest with Genetic Algorithm (GA-RF) gives the best performance of 87.42%.

Keywords: Machine Learning, Liver Disease Prediction, Genetic Algorithm

1 Introduction

The unhealthy lifestyle of people has lead to several new diseases which can become very dangerous if not diagnosed on time. Early diagnosis can be very beneficial thus machine learning becomes helpful as it can predict the diseases accurately and can save a lot of time. We focus on liver disease prediction in our work.

Liver disease is a disturbance in the typical functioning of the liver. Any pathological disorder in the liver can be a cause of liver disease. They have emerged as a prime cause of death worldwide, holding a position of twelfth by 2020. They cause around two million passing each year worldwide, out of which one million occur because of complications caused by cirrhosis and the rest one million are the consequences of hepatocellular carcinoma and viral hepatitis [1]. In 2010, Egypt experienced the largest age-standardized cirrhosis mortality rates - accounting for 72.7 deaths per 100,000 [2]. There are many variants of liver diseases found, but Hepatitis is the most common condition of them all. A, B, C, D and E are the major strains of Hepatitis which cause liver damage. A study conducted by WHO stated that out of the 325 million affected by Hepatitis B and C worldwide, 4.5 million demises would have been prevented through appropriate vaccination and effective diagnostic tests [3].

In the year 2018, World Health Organization (WHO) found that India had about 2, 64, 193 deaths caused by liver diseases [1]. This amounts to about 3% of the reported deaths worldwide. There are multiple causes of the sudden outburst of liver diseases in India, but alcoholism is the primary concern. Indians consumed 5.4 billion liters of alcohol in the year 2016 alone, and researchers projected that it would increase to 6.5 billion liters a year by 2020 [4]. These numbers highly concern and largely contribute to liver ailments in India. Obesity, diabetes, drug use, and other factor contribute to liver inflammation which ultimately may result in liver diseases. Researchers fear India will emerge as the world capital of liver diseases by the year 2025 [4]. With passaging days, the amount of patient medical records is continuously increasing. The usage of machine learning is hence becoming a driving force in the medical industry. Doctors can make an early diagnosis of the diseases using these machine learning based systems. These models even assist the doctors in making the correct treatment choices for the patient, and thereby, the large patient queue is immensely minimized at liver specialists.

In this paper, we experiment with a wide range of machine learning algorithms namely Multilayer Perceptron, Neural Networks, Random Forest, Support vector machine, and Bayesian Classifier to predict liver disease. Furthermore, genetic algorithms select a subset of features which improves performance of the models. Thereafter, we have compared the results obtained through these analyses on various parameters like accuracy, precision, and recall. These comparisons can help us deduce an effective solution to the problem.

2 Literature Review

This section contains the details of the previous works concerning prediction of liver diseases utilizing various machine learning algorithms. Table 1 provides information about some of the related papers, their authors, brief description about different ML algorithms used in these papers and results obtained for different ML algorithms.

Table 1: Comparison of research work with the same dataset

Author(s)	Brief Description	Results
Maria Alex Kuzhippalli et al.	In this paper, data is processed using data exploration, pre-processing like label encoders, replacement of null values with the median values, and outliers removal. Feature selection is performed on the processed data using the genetic algorithm, which follows initialization, selection, crossovers, etc., for finding the valuable features. Multilayer Perceptron, KNN, Logistic Regression, Decision Tree, Random Forest Tree,Gradient Boosting, AdaBoost, XGBoost, Light GBM, Stacking Estimator were used here.	Multilayer Perceptron:82% KNN:79% Logistic Regression: 76% Decision Tree:84% Random Forest Tree:88% Gradient Boosting:84%
Mafazalyaqeen Hassoon et al.	The genetic algorithm involves selection, crossover, and mutation, and this is used for data mining optimization and for finding attributes with the highest accuracy, which are extracted from boosted c5 algorithm. The data needs to be encoded for performing this, fitness functions are calculated, parents combined in crossover to produce children. The optimized rules are compared to the results with boosted c5.during implementation, datasets are divided between training and testing datasets, respectively. The; datasets train datasets are the ones on which optimizations are performed and compared to the test data.	Genetic Algorithm: 92.93% Boosted c5.0: 81.87%
C. Geetha et al.	It recognized the need to develop machine learning techniques to detect liver disease in India and used the data collected from already diagnosed patients. The two main algorithms proposed were SVM and Logistic Regression, which then resulted in the probability of liver disease prediction attained with an accuracy of about 96\%. Other techniques like the random forest, Naive byes classification, etc., were proposed as well, and bio-inspired optimization algorithms concentrated focus on the implementation of parametric classification.	Logistic Regression - 73.23% SVM - 75.04%

A.K.M. Sazzadur Rahman et al	This paper utilized six algorithms - Logistic Regression, K Nearest Neighbors, Decision Tree, Support Vector Machine, Random Forest, and Naive Bayes for liver disease diagnosis. We used various measurement techniques such as accuracy, precision, recall, f-1 score, and specificity to determine the performance of the trained machine learning models.	LR - 75% NB - 53% KNN - 62% Random Forest - 74% SVM - 64% Decision Trees - 69%
G. Shobana et al.	This paper stresses the need for the early detection of liver disease. Machine learning techniques can facilitate this early detection of diseases. Gradient Boost, XGBoost, LGBM Boost, and CatBoost are proposed and used with algorithms like Logistic regression, Linear discriminant Analysis, Naive Bayes, Decision Tree, etc. With the dataset provided, preprocessing with scaling is done. After training the models, the best-suited algorithm with the method is highlighted.	Cat Boost: 92.6% LGBM Classifier: 93.4% XGBoost: 91.9% Gradient Boosting: 91.5%
Jagdeep Singh et al.	In this paper, they have used classification and feature selection techniques to predict liver disease. The dataset used in this research paper to predict liver disease risk levels contains a lot of attributes likeage, gender, sgpt, direct bilirubin, total bilirubin, albumin, etc. Various Classification algorithms have been implemented on the liver patient dataset in this research paper like Naive Bayes, Logistic Regression, J48, Random forest algorithm, k-nearest neighbor to find the accuracy. Comparison between the accuracy of different classification algorithms is made with and without using feature selection technique.	Without using feature selection techniques: Logistic Regression - 72.50%, Naive Bayes - 55.74%, SMO - 71.35%, IBk - 64.15%, J48 - 66.78%, Random Forest - 71.53% Using feature selection techniques: Logistic Regression - 74.36%, Naive Bayes - 55.9%, SMO - 71.36%, IBk - 67.41%, J48 - 70.67%, Random Forest - 71.87%
Shapla Rani Ghosh et al.	In this paper, they have used certain selective algorithms on medical instruments (e.g., ECG, CT Scanner, Ultra Sono, MRI, etc.) to predict liver disease so as to lessen time and cost on hepatic disease diagnosis. This research paper's algorithms are KStar, Naive Bayes Classification (NBC), Logistic, Bagging, and REP tree to predict sensitivity, precision, accuracy, and specificity. Two datasets were used in this research paper, one from UCLA and another one from AP, to find out the best algorithm for liver disease diagnosis. The software used for this analysis was Weka 3.6.10. It has been predicted in this research paper that the KStar algorithm has the best precision, sensitivity, accuracy, and specificity, while minimum accuracy was obtained from NBC Algorithm. Hence this paper predicts that the KStar algorithm is suitable for the rapid identification of liver disorders.	Bagging - 84%, K- Star - 98.5%, Naive Bayes - 35%, Logistic - 75.6%, REPTree - 80.4%

Somaya Hashema et al.	The paper states HCC is a malignant tumor of the liver, and people with HCV are at higher chances to develop it. They can be detected accurately and efficiently with the help of machine learning. The adtrees model, CART and REPtrees, and the multi-linear regression were used on the dataset on five attribute names: age, AFP, alp, albumin, and total bilirubin.	Linear Regression- 96% Adtress - 99% CART - 95.5% Reptrees - 95.5%
Musavir Hassan et al.	This paper has used machine learning techniques like Artificial Neural Networks(ANNs) and Logistic Regression to predict diabetes disease at an early stage. In this paper, a general framework has been established to explain the functioning of Artificial Neural Networks in binomial classification and implement and evaluate the variants of the Back Propagation Algorithm. The dataset used in this research paper is the PIMA Indian	Logistic Regression - 76.8% Standard Back Propagation – 78.60% Resilient Back Propagation - 80.35% Variable Learning Rate - 77.73% Powell-Beale Conjugate Gradient - 79.04%, Levenberg Marquardt - 80.35% Quasi Newton Algorithm – 80.35% Scaled Conjugate Gradient – 79.84%

The researchers have attempted several times to find an effective method to utilize ML for liver disease prediction. Thirunavukkarasu et al. [5] used the following machine learning models – K- Nearest Neighbor (KNN), SVM and Logistic Regression (LR) to predict liver disease. LR and KNN models obtained a prediction accuracy of 73.97% each, whereas SVM had 71.97% prediction accuracy. They concluded that KNN and LR had higher accuracies. Vyshali J Gogi et al. [6] categorized the data using SVM, Linear Discriminant, LR and Decision Tree. They used the Liver Function Test to obtain the attributes.

Javad Hassannataj Joloudaria et al. [7] proposed a significant feature selection method by differentiating various data mining models for prediction of liver disease based on Extraction, Loading, Transformation, Analysis (ELTA) approach for appropriate diagnosis. Hence, the comparison of Random Forest, SVM, Bayesian networks, and MLP algorithms was done. It gave the major focus to the usage of clinical data for liver disease prediction and explored a variety of data through the analysis techniques used.

3 Proposed Methodology

This section provides information about the dataset through tables and heat maps, different steps followed during data pre-processing, feature selection techniques and at last provides information about different algorithms used on the dataset.

In this work we use the Indian Liver Patient dataset (ILPD) that has been obtained from UCI repository. The Data Mining process starts with a preprocessing step; after that feature selection is applied using the filter method. Different models are applied for the prediction of data using data mining techniques. Models like Random Forest, Bayesian Classifier, MLP Neural Networks, SVM, have been applied and we have also used hyperparameter tuning with the machine learning models to find an optimal combination of hyperparameters. Genetic Algorithm (GA) approach has been used to optimize Random Forest, SVM and Bayesian Classifiers and their performance has been evaluated using Confusion Matrix. Figure 1 illustrates the proposed methodology used in this experiment

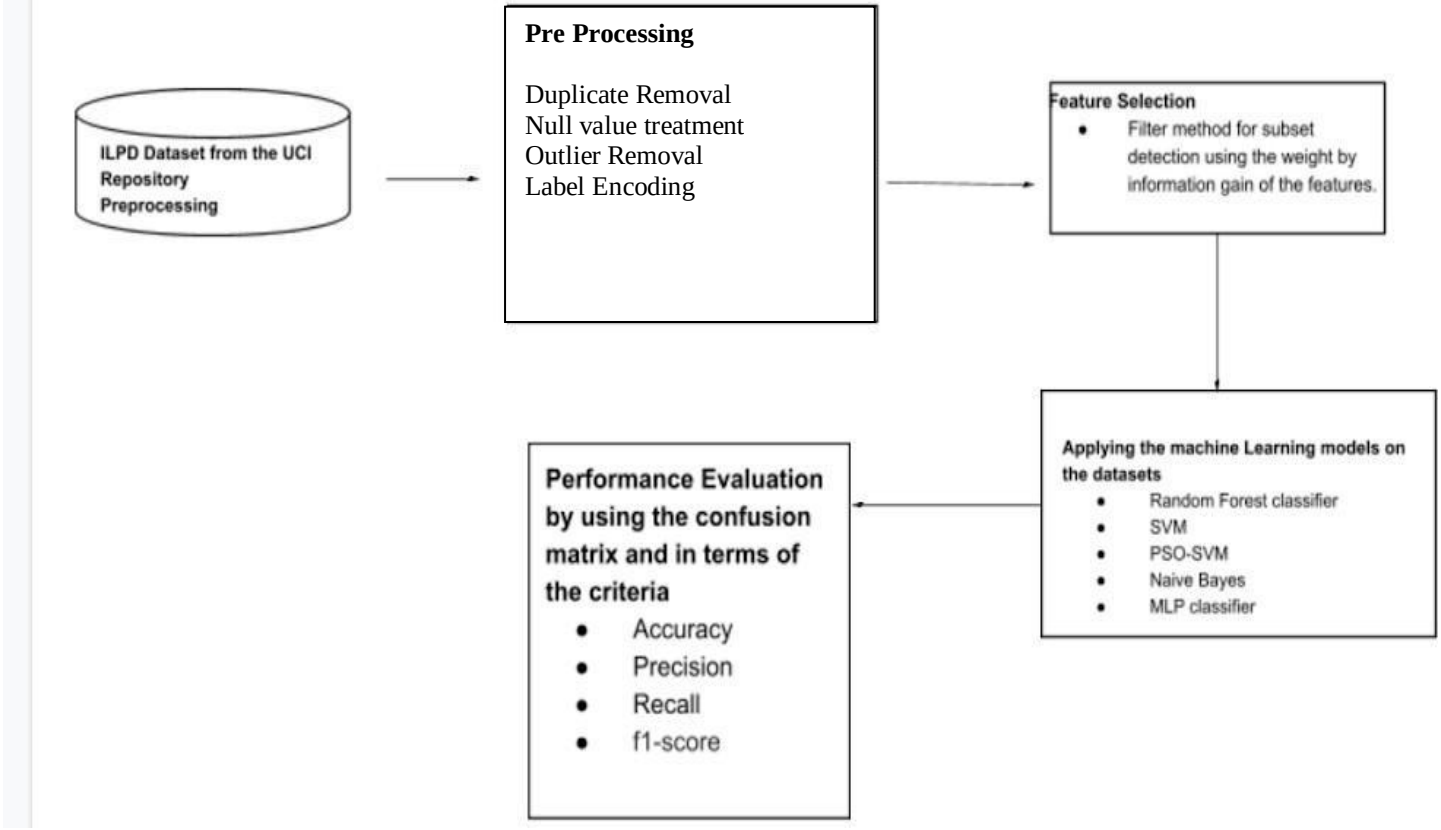

Figure 1: Proposed Methodology

3.1 Dataset Description

This section dicusses about the dataset and its different features. Table 2 provides information about different attributes of the dataset, its datatype, mean value of that column and range within which the value of that column lies. Table 3 tells about total no. entries, no, of patients with liver disease, no. of patients with no liver disease, no. of male patients, no. of male female patients. The details are as follows:

Table 2: Dataset features description table

Feature Name	Type	Mean	Range
Age	Real Number	44.75	4 to 90
Gender	Categorical	----	Male, female
Total Bilirubin (TB)	Real Number	3.3	0.4 to 75
Direct Bilirubin (DB)	Real Number	1.49	0.1 to 19.7
Alkaline Phosphate (Alkphos)	Integer	290.58	63 to 2110
Aminotransferase (Sgpt)	Integer	80.71	10 to 2000
Aminotransferase (Sgot)	Integer	109.91	10 to 4929
Total Proteins (TP)	Real Number	6.48	2.7 to 9.6
Albumin (ALB)	Real Number	3.14	0.9 to 5.5
Albumin and Globulin ratio (A/G Ratio)	Real Number	0.95	0.3 to 2.8
Target Class	Categorical	----	1 and 2

Table 3: Details of the base dataset

Details	Count
Total Entries	583
Patients with Liver disease	416
Patients with No liver disease	167
Male Patients	441
Female Patients	142

The correlation between the attributes mentioned above is given in the heatmap given below (Figure 2) which is obtained using the **corr()** function.

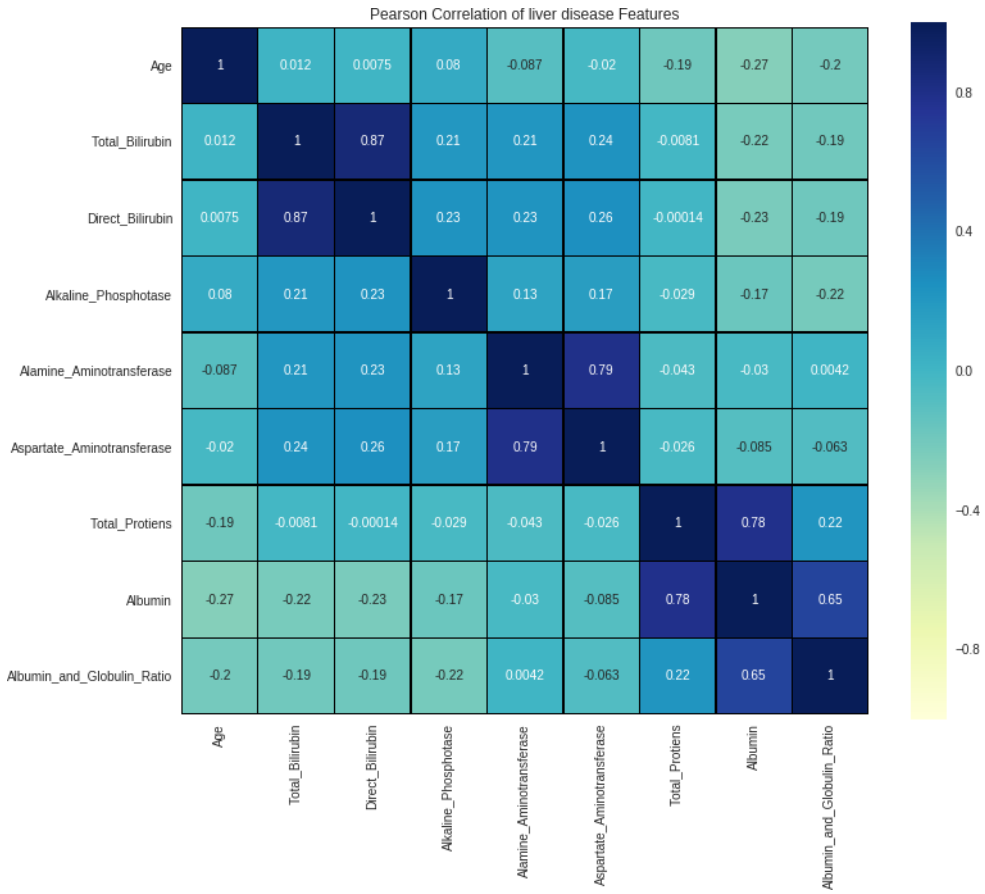

Figure 2: Heatmap of attributes present in Indian Liver Patients Dataset by UCI

3.2 Data Preprocessing:

For data preprocessing, we started with replacing the null values of the dataset with the mean values of that attribute e.g.: '*Albumin and the Globulin ratio'* had four null values, which were replaced with mean values. We removed 13 duplicate tuples and also removed the outliers in '*Aspartate Aminotransferase'* attribute. We have also performed Label Encoding of the Gender attribute.

The target data has two groups – 1 is the liver patients and 2 is the non-liver patients.

The dataset has been divided into 2 sets: training set and testing set. 60% of the dataset (350 records) were considered for training while the rest 40% (233 records) were considered for testing.

3.3 Feature Selection:

It is an essential step in data preprocessing. It involves identifying the redundant features and removing them from the relevant data parts through identification of the crucial features, thus helping to improve the machine learning models' accuracies. Filter-based feature selection method has been used for selecting the subsets using the weight by information gain of the features.

3.4 Model Construction:

It is essential to apply data mining models in medical research since large volumes of data is available. The following algorithms were used in this research paper.

3.4.1 Random Forest Classifier:

This is a supervised learning algorithm primarily used for classification which consists of many decision trees represented as n estimators and its accuracy was improved using hyperparameter tuning using GridSearchCV. Accuracy of the Random Forest Model came out to be 67.98% and after hypertuning accuracy came out to be 69.29%.

3.4.2 Bayesian Classifier:

It is one of the most simple and effective classification algorithms. Accuracy of the Bayesian network came out to be 55%. In this algorithm weights were assigned to each feature based on their probabilities of occurrence. It selects greatest probability features.

3.4.3 MLP Neural Networks:

MLP neural network output consists of three layers: input, middle and output layers. The accuracy of the model came out to be 65.79% and this is improved using hyperparameter tuning using the GridSearchCV. Accuracy obtained after performing hyperparameter tuning came out to be 71%.

3.4.4 Support vector machine:

It is a Supervised algorithm wherein each data item is plotted into an n-dimensional space, where the number of features present in the model are denoted by n. Here, the classification is done by finding a hyperplane that differentiates the various classes. The accuracy of the algorithm came as 69.74% with the hyperparameter tuning done with the help of GridSearchCV and the best values of C=1, gamma='scale', kernel='linear'.

3.4.5 Particle Swarm Optimization

It is used for the optimization of given data. It solves computationally complex optimization problems and is a population-based stochastic algorithm. The nature of birds inspires it. When the birds find the best source of food, they flock around it. The bird nearest to it leads the others towards the direction of the source. For the algorithm, the particle is considered to be an object with specific properties. So, here the particle is the best solution, and the various data objects swarm around it when found. The iterations are carried out so that the closeness to the optimal solution reaches.

3.4.6 Genetic Algorithm

Genetic Algorithms are based on the concepts of genetics and natural selection where we consider a population of possible solutions to a given problem. Recombination and mutation are performed on all these possible solutions, which then produce children. It is repeated over several generations. The fitness value is assigned to each of these children. These fitter children are provided with a greater probability to mate and produce ‘fitter’ children. This process is repeated until we reach the optimal solution.

3.4.7 Hyperparameter Tuning

It is used for selecting an optimal architecture for our machine learning models. So, we use a range of parameters called hyperparameters. Some parameters out of this help to create the most effective machine learning model architecture, and this process is called hyperparameter tuning.

We are using the GridSearch technique for hyperparameter tuning our machine learning models. It creates a multi-dimensional grid for all the possible values of these hyperparameters, and the best hyperparameter setting values are selected.

3.4.8 Genetic Algorithm with RF

Genetic algorithm is essentially used as a heuristic search tool which when combined with RF, which is a predicting tool, gives a model with improved accuracy than when only RF was used. Being a predictive model, RF will test the fitness of each individual. Further, the Genetic Algorithm will work by choosing the fittest individual in the entire population for the purpose of reproduction.

4 Performance Evaluation

We have used a confusion matrix for evaluating and analyzing the models. The confusion matrix will provide us four

values: True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN). We have represented these four values in the form of a table 4 given below for different machine learning models. These four values have also been used to evaluate sensitivity, specificity, accuracy, F-measure, precision, etc. The testing dataset has been used for calculating the accuracy of models.

Table 4: FN, TP, FP and TN values of models used

S.NO.	Machine Learning Model	TP	TN	FP	FN
1	Random Forest Classifier	150	7	7	60
2	SVM	159	42	28	67
3	MLP Neural Network	140	21	17	46
4	Naïve Bayes	76	64	83	3
5	GA - Random Forest classifier	147	19	52	15
6	GA - SVM	81	19	23	42
7	GA - Naïve Bayes	76	3	83	64

$$\text{Specificity} = \text{TNR} = \frac{TN}{TN + FP} \quad \ldots (1)$$

$$\text{Sensitivity or Recall} = \text{TPR} = \frac{TP}{TP + FN} \quad \ldots (2)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad \ldots (3)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad \ldots (4)$$

$$\text{FPR} = 1\text{- Specificity} \quad \ldots (5)$$

Figure 3: Equations used for Performance Evaluation of Machine Learning Models

5 Results and discussions

Performances of different models are compared by the help of results obtained from confusion matrix represented in Table 5. Accuracies of different models is represented in Table 6. On comparison, we found that the accuracy for the random forest classifier came out to be as 70%, accuracy for Bayesian Classifier came out to be 61%, accuracy for MLP Neural Networks came out to be 72%, and accuracy for SVM came out to be 70.35%.But with the addition of the Genetic Algorithm, the accuracies turned out to be 87.42% for Random Forest classifier, 73.06% for SVM and 71% for Naïve Bayes. Hence, through the differentiation of the model's accuracies we can conclude that after applying feature selection MLP neural network gave the best accuracy and Random Forest gave the best accuracy when we paired it with Genetic Algorithm.

Table 5: Confusion matrix of Machine Learning Models

S.NO.	Machine Learning Model	Specificity	Sensitivity	Precision	FPR
1	Random Forest classifier	0.5	0.714	0.955	0.5
2	SVM	0.6	0.70	0.85	0.4
3	MLP Neural Network	0.55	0.75	0.89	0.45
4	Naïve Bayes	0.43	0.96	0.74	0.57
5	GA - Random Forest classifier	0.44	0.73	0.88	0.55
6	GA - SVM	0.68	0.77	0.81	0.31
7	GA - Naïve Bayes	0.48	0.95	0.96	0.44

Table 6: Accuracy of Machine Learning models

S.NO.	Machine Learning Model	Accuracy
1	GA - Random Forest classifier	0.8742
2	GA - SVM	0.7306
3	GA - Naïve Bayes	0.71
4	Random Forest classifier	0.70
5	SVM	0.7035
6	MLP Neural Network	0.72
7	Naïve Bayes	0.61

6 Conclusion and future works

In recent times, it is essential to find a systematic approach for liver disease prediction. There were multiple data mining methods performed on the Liver disease dataset provided by UCI. This paper is primarily focused on finding the most appropriate method for achieving the best accuracy using the ELTA approach. Various algorithms including SVM, MLP Neural Network, PSO-SVM, Random Forest and Bayesian Network have been used. The authors have also performed feature selection and utilized the genetic algorithm to improve the accuracy. Furthermore, the authors have compared the models regarding their accuracy, precision, specificity, and sensitivity.

Consequently, these findings can aid the prediction of liver disease among individuals. Many other algorithms such as the Ant Colony System (ACS) can be used to further optimize the machine learning algorithms used in the paper. Also, the application of smart algorithms and deep neural networks in the interpretation and prediction of liver diseases, as well as using larger and more diverse datasets, further enhancement of accuracy can be obtained.

References

[1] Asrani SK, Devarbhavi H, Eaton J, Kamath PS. Burden of liver diseases in the world. J Hepatol. 2019 Jan;70(1):151-171. doi: 10.1016/j.jhep.2018.09.014. Epub 2018 Sep 26. PMID: 30266282.Ali A Mokdad, Alan D Lopez, Saied Shahraz et al.

[2] Mokdad AA, Lopez AD, Shahraz S, Lozano R, Mokdad AH, Stanaway J, Murray CJ, Naghavi M. Liver cirrhosis mortality in 187 countries between 1980 and 2010: a systematic analysis. BMC Med. 2014 Sep 18;12:145. doi: 10.1186/s12916-014-0145-y. PMID: 25242656; PMCID: PMC4169640.

[3] WHO
Hepatitis
https://www.who.int/health-topics/hepatitis#tab=tab_1

[4] India may become 'world capital of liver diseases'
Ummid, Saturday April 19, 2014 https://ummid.com/news/2014/April/19.04.2014/india-capital-liver-disease.html

[5] K. Thirunavukkarasu, A. S. Singh, M. Irfan and A. Chowdhury, "Prediction of Liver Disease using Classification Algorithms," 2018 4th International Conference on Computing Communication and Automation (ICCCA), 2018, pp. 1-3, doi: 10.1109/CCAA.2018.8777655.

[6] V. J. Gogi and V. M.N., "Prognosis of Liver Disease: Using Machine Learning Algorithms," 2018 International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE), 2018, pp. 875-879, doi: 10.1109/ICRIEECE44171.2018.9008482

[7] Javad Hassannataj Joloudari, Hamid Saadatfar, Abdollah Dehzangi, Shahaboddin Shamshirband, Computer-aided decision-making for predicting liver disease using PSO- based optimized SVM with feature selection, Informatics in Medicine Unlocked, Volume 17, 2019, 100255, ISSN 2352-9148,

[8] Dua, D. and Graff, C.
UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]
Irvine, CA: University of California, School of Information and Computer Science

[9] Singh, A., Singh, N. An approach for predicting missing links in social network using node attribute and path information. Int J Syst Assur Eng Manag (2021). https://doi.org/10.1007/s13198-021-01371-w

[10] Ankita and N. Singh, "A Link Prediction Model using Similarity and Centrality based Features," 2019 6th International Conference on Computing for Sustainable Global Development (INDIACom), 2019, pp. 415-417.

[11] Singh AN (2019a) Improved link prediction using PCA. Int J Anal Appl 17(4):578–585