IJARIIT INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078

(Volume 9, Issue 5 - V9I5-1136)
Available online at: https://www.ijariit.com

Liver disease detection using Machine Learning techniques

Taranpreet Kaur
taranpreet67670@gmail.com
DAV Institute of Engineering and Technology, Jalandhar,Punjab

Dr. Vinay Chopra
vinaychopra222@yahoo.co.in
DAV Institute of Engineering and Technology, Jalandhar,Punjab

ABSTRACT

Liver disease is the leading reason of death worldwide , The liver is responsible for metabolic, strength-storing, and waste-filtering functioning in your body. The aim of this study is to developing a machine learning based technique for liver disease prediction in people. This study on liver disease detection models meant to determine the best techniques for selecting and synthesising the many studies of high quality. The majority of health data is nonlinear, correlation-structured, and complex, make it complex to evaluate. The use of ML based techniques in healthcare has been ruled out. In this work use various machine learning algorithim like decision tree, Naïve Bayes, SVM , Random Forest , CatBoost ,Soft Voting Classifier on Indian Liver patient dataset to predict liver disease. The research work gives the correct or maximum accuracy model show that the model is able of predicting liver diseases effectively. Our end result shows that Voting classifier attain the higher accuracy as compared to other machine learning models.

Keywords: *Liver Disease, SVM, NB, Random forest, Catboost, Soft voting classifier*

I. INTRODUCTION

The liver is the largest organ in your body. Untreated liver infections can lead to liver failure and malignancy. Liver disease is the leading cause of death worldwide .According to the World Health Organization around 58 million people having hepatitis C infection .Hepatitis C was an main cause of death for around 290,000 people in 2019 ,mainly due to the cancer of the liver and scarring.

The liver is responsible for metabolic, strength-storing, and waste-filtering functioning in your body. The liver helps to eliminating hazardous compounds from the body, such as drugs and alcohol. It may also store many types of vital elements, such as vitamins, minerals, and glucose, and transport them into circulation as needed. Cirrhosis and hepatitis kill around 2 million people worldwide.

ML is widely used in healthcare industry for analyse various types of disease. When someone is currently in bad health they must schedule an expensive and time-consuming doctor appointment. Also, it can be not easy for the user if they are far from health facilities because the condition cannot be recognized. So, it can be better for the patient and make the process run more smoothly if the aforementioned operation can be carried out utilising an automated software that saves time and money .

The aim of this study seeks to to conquer these challenges by developing a machine learning based technique for liver disease prediction in people. Jaundice, liver enlargement, swelling in the legs and ankles, itchy skin ,vomiting ,dark urine color, weight loss and weakness in the muscles are some of the typical liver symptoms.

Inflammation is the initial stage of liver disease. The liver helps to eliminate toxic waste in our body. When the liver is unable to handle this toxic waste ,the body respond by enlarging the liver. If the swelling is not treated properly ,It can lead to scarring.

This stage is called fibrosis and at that stage your liver does not work properly. Cirrhosis is the third stage of liver disease .At Cirrhosis stage liver become very scarred and area around the liver itches. Liver failure is the final stage of liver disease .In this stage of liver failure ,the liver loses its ability to function and changes of liver cancer is increasing at this stage.

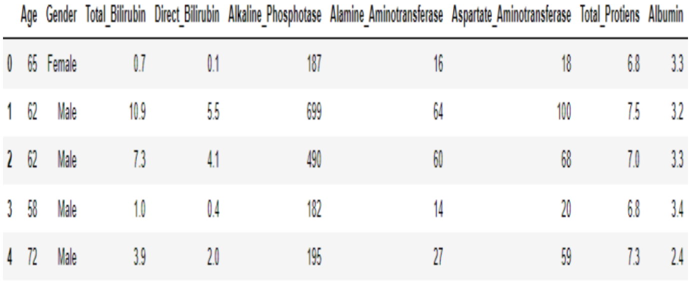
II. METHODOLOGY

2.1 DATA COLLECTION

The Indian Liver Patient Records dataset was used for the prediction of liver disease in this research .The aim of the dataset is to verify whether a person has liver disease or not. The dataset contain 583 rows and 11 columns, with one output column that indicates whether the person has been diagnosed with liver disease or not. There are 441 male and 142 female patient records in the given dataset.

S.NO	Column Name
	Age
	Gender
	Total Bilirubin
	Direct Bilirubin
	Alkaline Phosphate
	Alamine Aminotransferase
	Aspartate Aminotransferase
	Total Protein
	Albumin
	Albumin and Globulin Ratio
11.	Output

Table 1: Liver Patient Record Dataset Columns



	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin
0	65	Female	0.7	0.1	187	16	18	6.8	3.3
1	62	Male	10.9	5.5	699	64	100	7.5	3.2
2	62	Male	7.3	4.1	490	60	68	7.0	3.3
3	58	Male	1.0	0.4	182	14	20	6.8	3.4
4	72	Male	3.9	2.0	195	27	59	7.3	2.4

Figure 1:Fetch the Liver Disease Dataset using Python

Figure 2 shows that total 71% population will get liver disease and 29% population are not diagnosed with liver disease**.**

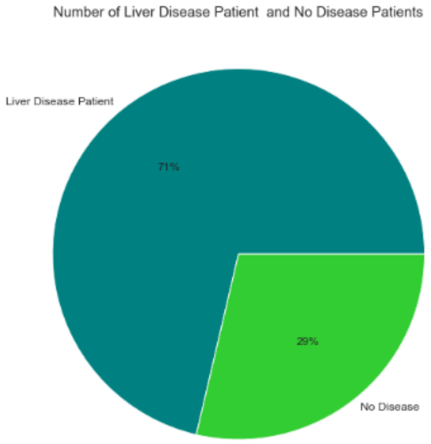
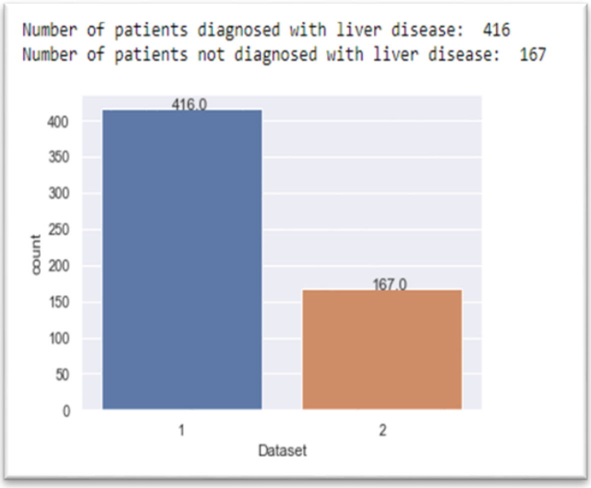


Figure 2 : Distribution of Liver Disease Patient and No Disease Patient

Figure 3 shows that out of 583 total 416 persons are diagnosed with liver disease and 167 are not diagnosed with liver disease.A status ‘1’ refers to a liver patient and status ‘2’ refers to a non liver patient.



```
Number of patients diagnosed with liver disease:  416
Number of patients not diagnosed with liver disease:  167
```

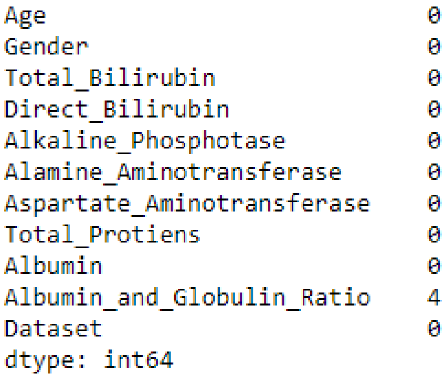
Figure 3 : Number of Patient Diagnosed with Liver Disease Patient and Not

2.2 Data Preprocessing

Data preprocessing is the method of preparing raw data for use with a machine learning model.It is the most required phase of machine learning in order to develop a most reliable and accurate ML model. Data pre-processing is required to cleaning the data. Data preparation is the process of eliminating the duplicate data, dealing with NAN values, and identified the outliers in a given dataset.

i. **Handling null and missing value**:In order to identified NAN values in the dataset we use the isnull() method .

Figure 4:illustrate that there is 4 null value present in the albumin and globulin ratio column in the liver disease dataset.We can replace the missing value with the mean.of the respective column.



```
Age                            0
Gender                         0
Total_Bilirubin                0
Direct_Bilirubin               0
Alkaline_Phosphotase           0
Alamine_Aminotransferase       0
Aspartate_Aminotransferase     0
Total_Protiens                 0
Albumin                        0
Albumin_and_Globulin_Ratio     4
Dataset                        0
dtype: int64
```

Figure 4: Observing the null and missing values

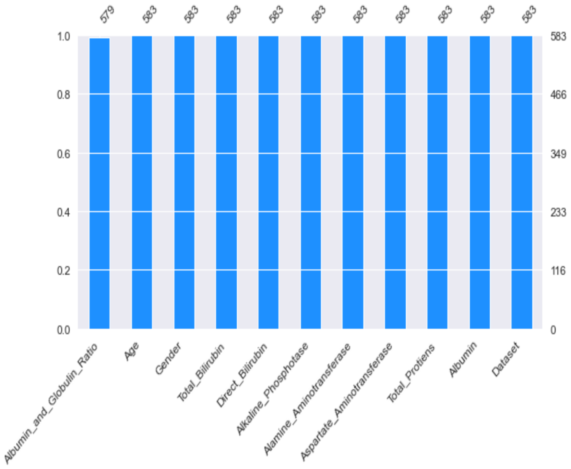
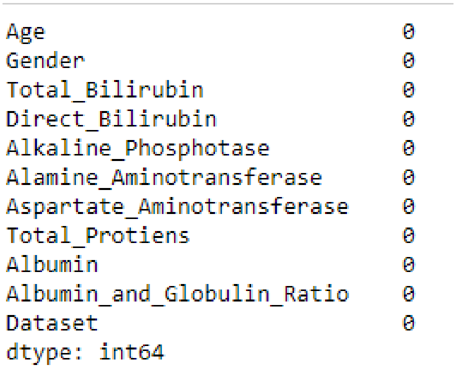


Figure 5 Visualizing the null values



```
Age                           0
Gender                        0
Total_Bilirubin               0
Direct_Bilirubin              0
Alkaline_Phosphotase          0
Alamine_Aminotransferase      0
Aspartate_Aminotransferase    0
Total_Protiens                0
Albumin                       0
Albumin_and_Globulin_Ratio    0
Dataset                       0
dtype: int64
```

Figure 6:Dataset after filling missing value

2.2.2 Handling the outliers: Outliers are those data points that are different from the other data points of the dataset. To identify the outliers box plot technique are used . From the **Figure 7** we clearly observe that age and albumin column do not have any outliers and rest all the columns contain have huge number of outliers. To eliminate the outliers qunatile transformer method is used in the given dataset.

:

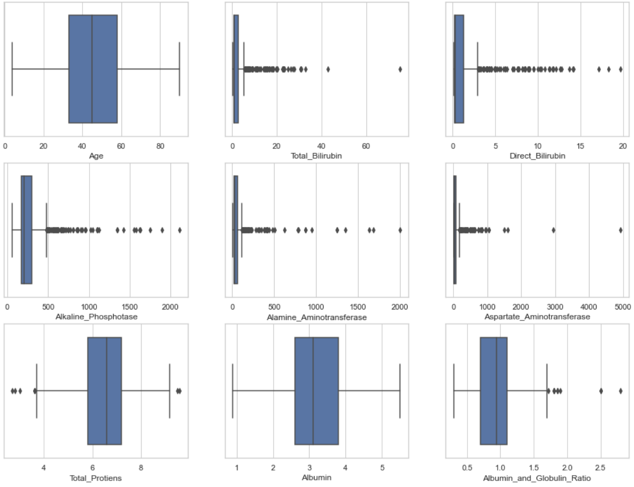


Figure 7: Observing the Outliers

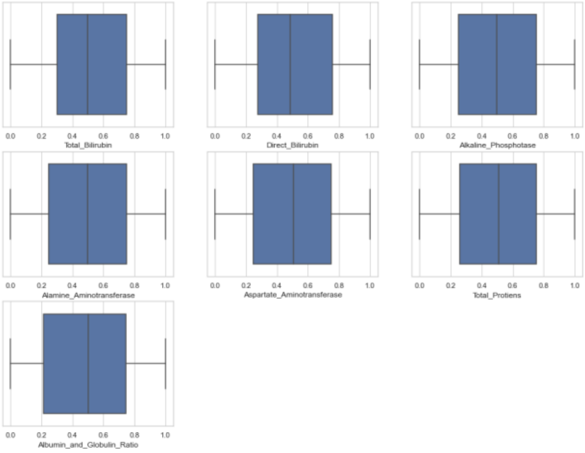


Figure 8:Removal of outliers using Qunatile Transformer Technique

2.2 Feature Engineering

It is the procedure of discovery the new quality and modify the existing features to get better the performance of machine learning model .It involve the selecting the relevant features ,transforming the variables ,dealing with null values. Pearson's correlation technique is a popular process to find the most significant attributes/features. . The correlation matrix is used to reveal the connection among the attributes and produces a matrix as an output. The correlation coefficient is calculate in this method, which correlate with the output and input attributes. The coefficient value remains in the range by between −1 and 1. The value above 0.5 and below −0.5 indicates a notable correlation, and the zero value means no correlation

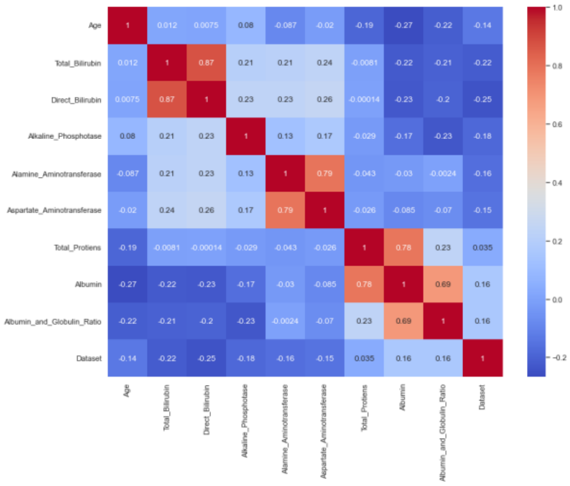
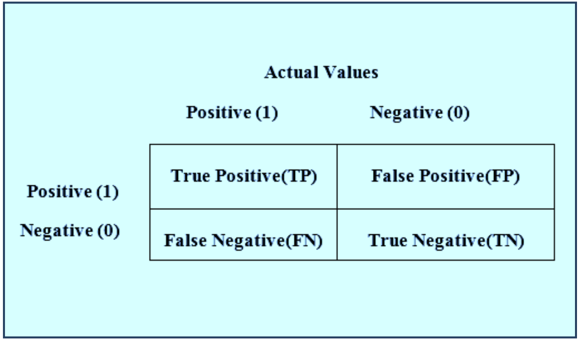


Figure 9:Correlation Matrix

2.3 Evaluation and Performance Measure Parameter

The performance matrices are used to discover the most accurate classifier for liver disease prediction. Confusion matrix which offers an output matrix that fully describes the model's performance displays the true positive, false positive, true negative and false negative by comparing the model prediction with actual outcomes. A true positive specify that person has disease, and the prediction also show a positive. A true negative show that person does not have a disease and the prediction also has a negative. False positive result demonstrate that person does not have the disease but the prediction is positive.False negative indicate that person having disease and the prediction is also positive.



	Actual Values	
	Positive (1)	Negative (0)
Positive (1)	True Positive(TP)	False Positive(FP)
Negative (0)	False Negative(FN)	True Negative(TN)

Figure 10:Confusion Matrix

We have chooses accuracy matrix, precision, recall and F1 value to calculate the performance of all the models.The accuracy is the ratio of number of correct prediction to the total number of predictions made. Precision is a performance matrices used in ML to evaluate quality of model predictions. Recall define as the out of total positive classes, how our model predict accurately. There must be higher possible recall **.** If two model have small precision and high recall or vice versa, it is complicated to compare these models. So, for this reason, we can utilize F-score. This score help us to calculate the recall and precision at the similar time. The F-score is greatest if the recall is same to the precision.

$$\text{Accuracy} = \frac{\text{Number of Correct prediction}}{\text{Total number of predictions made}}$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

$$\text{F1 Score} = \frac{2 * (\text{Precison} * \text{Recall})}{(\text{Precision} + \text{Recall})}$$

2.4 Classification Techniques

2.4.1 Support vector machine:

SVM is the most commonly used ML algorithim .It is helpful in classification and regression problems.It is used to recognize the best separation hyperplane between classes by locating the set of points on the class descriptors' edges. The margin is the distance between the classes. SVM algorithms find a margin with the greatest possible distance. The greater the margin, the higher the classification accuracy of the classifier. The data points on the boundary are referred to as support vectors.

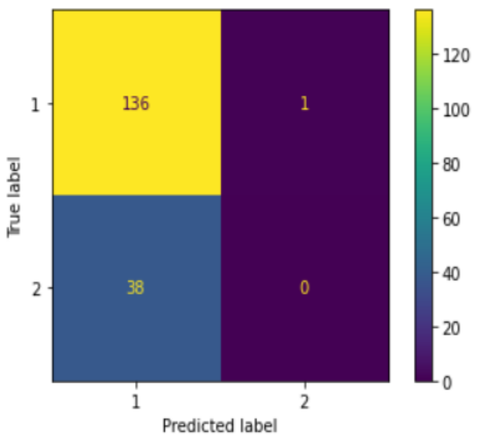


Figure 11:Confusion matrix of SVM

2.4.2 Logistic Regression:

For binary classification issues, where the objective is to predict a binary output (either 0 or 1) based on a set of input features, one type of ML approach is logistic regression. Its simplicity and effectiveness make it a popular approach in the fields of statistics and machine learning.In logistic regression, the connection between the input features and the output variable is model by a linear equation. The logistic function, also referred to as a sigmoid function, is used to convert the output of the linear equation into a probability between 0 and 1. The anticipated probability of the positive class (i.e., class 1) is then interpreted from the probability output.

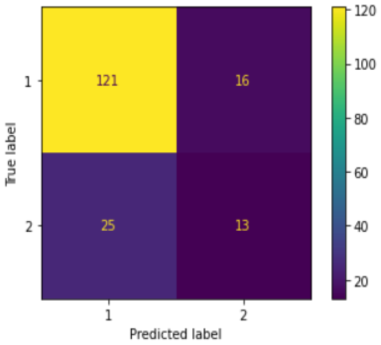


Figure 12:Confusion matrix of Logistic Regression

2.4.3 NAIVE BAYES CLASSIFIER

NB algorithm is a supervised learning algorithm, which is based on bayes theorem and used for solving classification problems. Being a probabilistic classifier, it make prediction based on the probability that an object will occur. Bayes' theorem is, which is used to find out the probability of a hypothesis with earlier knowledge. It is based on on the conditional probability.

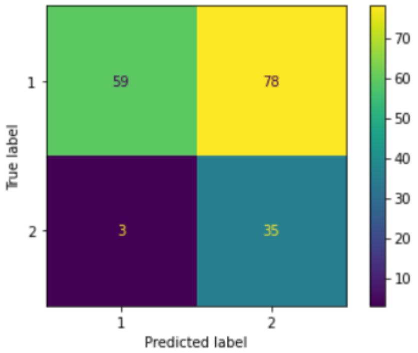


Figure 13:Confusion matrix of Naïve Bayes

2.4.4 Random Forest

It is a set of classification-based trees that have not been trimmed It performs incredibly well with regard to a variety of real-world issue because of its inconsiderateness to dataset noise and incredibly low danger of overfitting. In association to several additional tree-based algorithms, it operates more quickly and typically enhance data accuracy for testing and validation. RF is created by combining the predictions of different decision tree algorithms. There is diverse choice for adjusting the random forest's act when building a random tree.

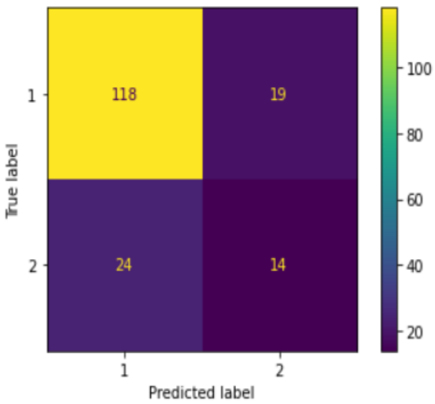


Figure 14:Confusion matrix of Random Forest

2.4.5 Soft Voting Classifier

Soft voting classifier is also known as weighted average ensemble classifier .Soft voting classifier is used in classification .A Weighted average classifier integrate the result of various classifiers in order to give a final result .It can also helpful for handling and managing the unbalanced dataset The soft voting classifier enhances the probability of identifying the most reliable and accurate classifier and produces the final correct prediction by giving weights to numerous classifiers.

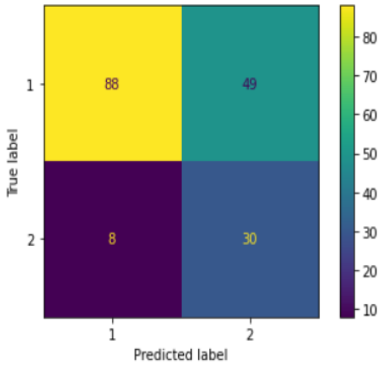


Figure 15:Confusion matrix of Voting Classifier

2.4.6 CatBoost Classifier

Catboost stands for categorical boosting. Catboost automatically handle the missing values during the training by using the symmetric weighted quantile sketch algorithim. It is a GBM classifier modification that can handle both category and numerical features.. It has the potential to improve classifier performance while decreasing overfitting and adjusting time. CatBoost uses parallel computing during the training process, which makes it faster than other gradient-boosting implementations.

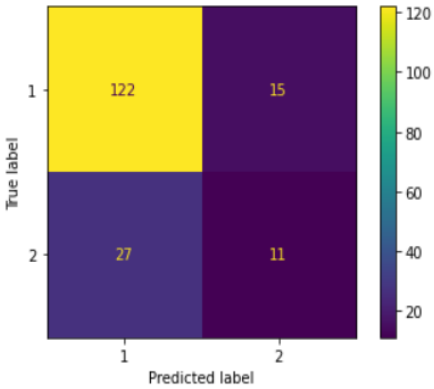


Figure 16:Confusion matrix of CatBoost

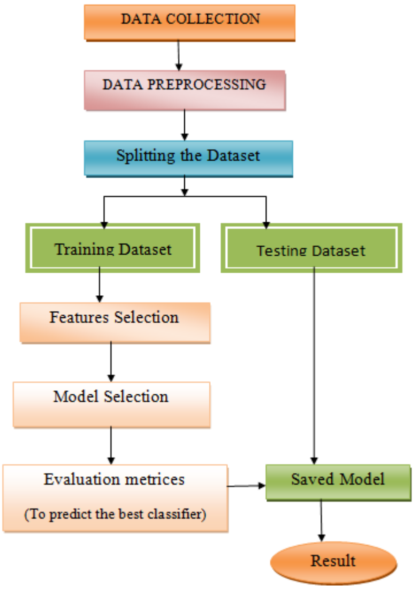


Figure 17: Overview of the Process

3. RESULT AND EXPERIMENT ANALYIS

The goal of this research is to predict the liver disease detection using various machine learning classifier.

Table 2:shows the results of the various ML classifiers utilizing all available features.The Logistic regression algorithim achieve 0.82 precision ,0.88 recall and 0.85 F1 score,while the random forest algorithim get the 0.82 precision ,0.86 recall and 0.85 F1 score value.SVM has accuracy on 0.78 precision,0.99 recall and 0.85 F1 value.Naive bayes classifier obtain the 0.90 precision,0.43 recall and 0.59 f1 value. The voting classifier attain the 0.91 precision ,0.65 recall and 0.76 F1 score value.The catboost attains the 0.82 preciison,0.82 recall,0.81 F1 score value.

ML Classifiers	Precision Value	Recall Value	F1-Score
Logistic Regression	0.82	0.88	0.85
Random Forest	0.82	0.86	0.84
SVM	0.78	0.99	0.85
Naive Bayes	0.90	0.43	0.59
Soft Voting Classifier	0.91	0.65	0.76
CatBoost	0.82	0.81	0.81

Table 2:Classification Performance of all Classifiers(Models)

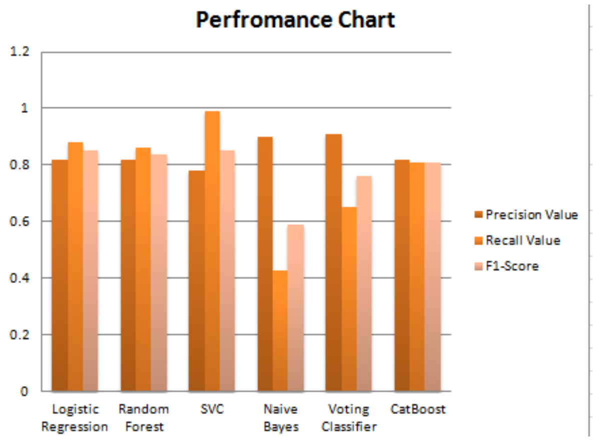


Figure 18 :Classification Performance Chart of all Classifiers(Models)

Table 3: We can see from the accuracy table that the Voting classifier perform the best and produces the most accurate results as compared to other machine learning classiifer, with an accuracy rate of 78%. With 77% and 76% accuracy, respectively, the SVM and Logistic regression classifiers outperform the Voting classifier**.**

ML Classifiers	Accuracy
Logistic Regression	0.76
Random Forest	0.75
SVC	0.77
Naïve bayes	0.53
Soft Voting Classifier	0.78
CatBoost Classiifer	0.71

Table 3:Accuracy Table of all Classifiers(Models)

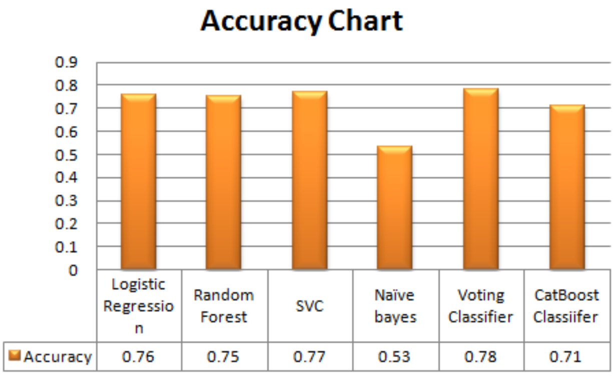


Figure 19: Accuracy Chart of all Classifiers(Models)

4. CONCLUSION

Liver Disease related deaths are on the rise. As a result, it is critical to develop a model and system that can easily predict liver disease at an early stage. The main goal was to find the best machine learning classifier for properly identifying liver disease.. This study compares Logistic Regression, SVM, Random Forest, Naïve Bayes, voting classifier and Catboost machine learning classifiers for liver disease detection using the Indian liver patient dataset. The voting classifier was discovered to be the most accurate in identifying liver disease, with a 78% accuracy score. The SVM outperforms all other algorithms after voting classifier , with an accuracy of 77%.

5. REFERENCES

[1]. Khan, Md Ashikur Rahman, Faria Afrin, Farida Siddiqi Prity, Ishtiaq Ahammad, Sharmin Fatema, Ratul Prosad, Mohammad Kamrul Hasan, and Main Uddin. "An effective approach for early liver disease prediction and sensitivity analysis." *Iran Journal of Computer Science* (2023): 1-19.

[2]. Amin, Ruhul, Rubia Yasmin, Sabba Ruhi, Md Habibur Rahman, and Md Shamim Reza. "Prediction of chronic liver disease patients using integrated projection based statistical feature extraction with machine learning algorithms." *Informatics in Medicine Unlocked* 36 (2023): 101155.

[3]. Shaheen, H., K. Ravikumar, N. Lakshmipathi Anantha, A. Uma Shankar Kumar, N. Jayapandian, and S. Kirubakaran. "An efficient classification of cirrhosis liver disease using hybrid convolutional neural network-capsule network." *Biomedical Signal Processing and Control* 80 (2023): 104152.

[4]. Nam, David, Julius Chapiro, Valerie Paradis, Tobias Paul Seraphin, and Jakob Nikolas Kather. "Artificial intelligence in liver diseases: Improving diagnostics, prognostics and response prediction." *JHEP Reports* 4, no. 4 (2022): 100443.

[5]. Srivastava, Aviral, V. Vineeth Kumar, T. R. Mahesh, and V. Vivek. "Automated Prediction of Liver Disease using Machine Learning (ML) Algorithms." In *2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, pp. 1-4. IEEE, 2022.

[6]. Cheng, Runtan, Lu Wang, Shenglong Le, Yifan Yang, Can Zhao, Xiangqi Zhang, Xin Yang et al. "A randomized controlled trial for response of microbiome network to exercise and diet intervention in patients with nonalcoholic fatty liver disease." *Nature Communications* 13, no. 1 (2022): 2555.

[7]. Sivasangari, A., Baddigam Jaya Krishna Reddy, Annamareddy Kiran, and P. Ajitha. "Diagnosis of liver disease using machine learning models." In *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, pp. 627-630. IEEE, 2020.

[8]. Choudhary, Ritesh, T. Gopalakrishnan, D. Ruby, A. Gayathri, Vishnu Srinivasa Murthy, and Rishabh Shekhar. "An Efficient Model for Predicting Liver Disease Using Machine Learning." Data Analytics in Bioinformatics: A Machine Learning Perspective (2021): 443-457.

[9]. Rabbi, Md Fazle, SM Mahedy Hasan, Arifa Islam Champa, Md AsifZaman, and Md Kamrul Hasan. "Prediction of liver disorders using machine learning algorithms: a comparative study." In 2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT), pp. 111-116. IEEE, 2020.

[10]. Wu, Chieh-Chen, Wen-Chun Yeh, Wen-Ding Hsu, Md Mohaimenul Islam, Phung Anh Alex Nguyen, Tahmina Nasrin Poly, Yao-Chin Wang, Hsuan-Chia Yang, and Yu-Chuan Jack Li. "Prediction of fatty liver disease using machine learning algorithms." Computer methods and programs in biomedicine 170 (2019):2311,-29.