Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7923033

LIVER DISEASESDIAGNOSIS AND LIVER DISEASE STAGE PREDICTION USING HYBRID MACHINE LEARNING CLASSIFIERS

Mr. Sagar Patel[1], Dr. Chintan Shah[2],Dr. Premal Patel[3]
PhD Scholar1 ,Associate Professor[2,3,]
Department of Computer Engineering[1,2,3]
College of Technology[1,2,3]
Sagarpatel.alpha@gmail.com[1], chintan.shah84@gmail.com[2],
premalpatel.ce@socet.edu.in[3]
Silver Oak University, Ahmedabad, Gujarat[1,2,3].

ABSTRACT

During the recentdecades, the risk of Liver disease in people is increasing at a rapid rate and is sought to beone of the fatal diseases in the world. It's quite a difficult task for researchers to predict the disease fromhumongous medical databases. To combat this issue, they have come up with machine learning techniques likeclassification and clustering. The main aim of this Research is to predict the chances of a patient having a liverdiseaseusingtheclassificationalgorithms. And it identify the stage of Liver disease like 1-Cirrhosis Liver, 2-Liver fibrosis, 3-Fatty Liver, 4-Healthy Liver. So NB, SVM, LOR,RF,DT,KNN, RBTC thesealgorithmsarecomparedwith proposed Hybrid Classifier(RF,SVC,XGBoost)basedontheirclassificationaccuracy and execution time. With these performance factors taken into consideration, the Hybrid Classifier whichserves as a better classifier is chosen with 99% accuracy.

Keywords:LogisticRegression,NeuralNetwork,Dataset,Accuracy,SVM, HYBRID model.

I. INTRODUCTION

This Research provides the software which facilitates to upload the details and get to know the prediction forLiverdisease. This Research uses Machine Learning algorithms to classify whether the liver condition is normal.We use NB, SVM, LOR,RF,DT,KNN, RBTC and Hybrid models for the prediction. This model will be useful for healthindustries who need to predict the diseases. The model will be helpful to know whether the liver condition isnormal or abnormal using the blood reports of the patient. This information regarding the patients will behelpfulforthemedicalcompaniesintheprocess.Theexistingmodelsincludevariousmachinelearningtechniqueswhichyieldoutputsoflessaccuracyandcan'thandlelargebundlesofdata.Thepoorperformancein the training and testing of the liver datasets is observed. These previously designed systems have beensufficient but more work has to be done on their prediction rate for better accuracy in the diagnosis of the liverdisease. The proposed system here uses concept of machine learning, and the models are first trained, thentested. Finally the most accurate model will predict the final result. Initially, the system asks you to enter yourdetails including age, gender, total Bilirubin, direct Bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT andAlkphos. These valuescan be known byblood test report of the user. After taking these inputsfrom the user,the system compares the data input with the training dataset of most accurate model and then predicts theresult accordingly as risk or no risk. The algorithms used are Logistic Regression, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest(RF), Decision Tree(DT), Naïve Bayes (NB), Hybrid Classifier(RF, SVC, XGBoost) etc. The dataset used is The IndianLiverPatient Dataset (ILPD) which was selected from UCI Machine learning repository. It is a sample of the entireIndian population collected from Andhra Pradesh region and comprises of 585 patients data. The system is verysimple in design and to implement. The system requires very low system resources and the system will work inalmostall configurations.

II. METHODOLOGY

Thevariousstagesinvolvedare:

ExploratoryDataAnalysis

Data visualization: With the help of data visualization, we can see how the data looks like and what kind ofcorrelation is held by the attributes of data. It is the fastest way to see if the features correspond to the outputfeatures.

Correlation Analysis: Correlations have three important characteristics. They can tell us about the direction ofthe relationship, the form (shape) of the relationship, and the degree (strength) of the relationship between twovariables.

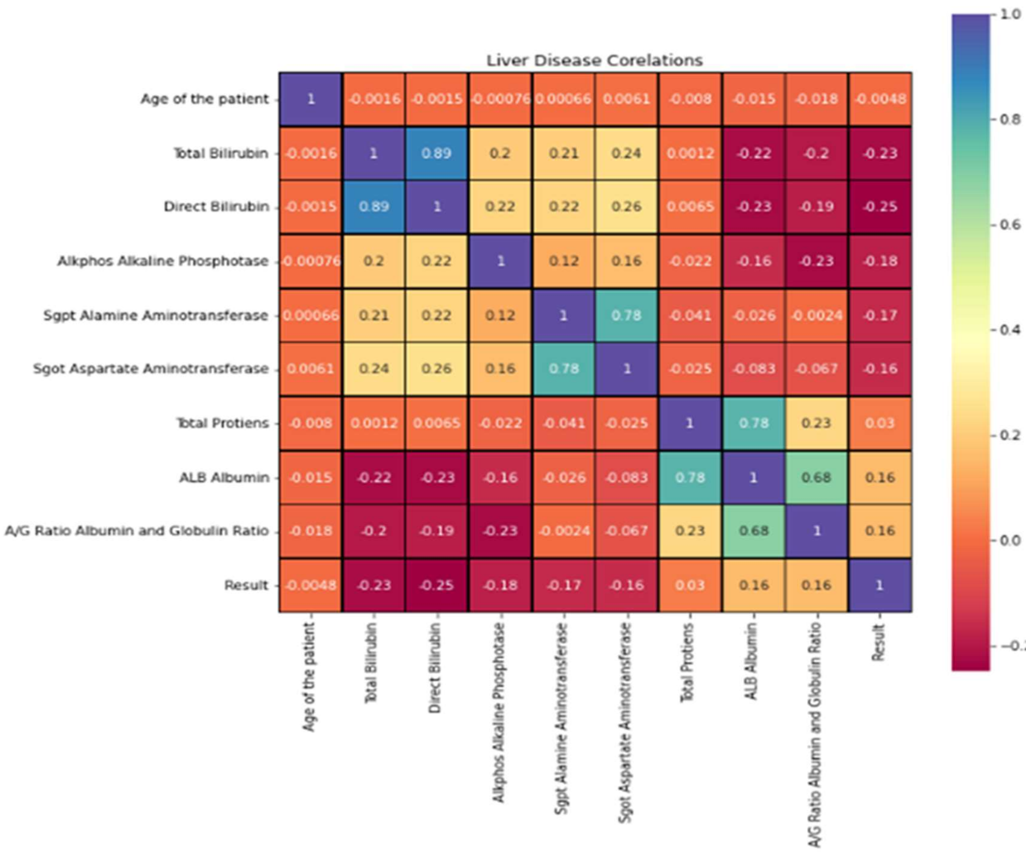


Figure1:CorrelationMatrixoftheModel

Data Preprocessing

This involves eliminating the null and most common words from the text. The words in the dataset consists oflinks, multiple full stops, very long and short words. These all need to be eliminated before providing it to thealgorithm. The significant stages in data preprocessing are Data Cleaning, Data Integration, Data Reduction andData Tranformation. It is carried out to meet the criteria of accuracy, completeness, consistency, timeliness,believabilityandinterpretability.

TrainingClassificationModel

We split the dataset into testing and training in multiple ratios to give the best results. Now we train the modelusing the Machine Learning algorithms namely: NB, Logistic Regression, RF, DT, KNN ,SVMand Hybrid Ensemble Classifier to predict the exactresult.

III. MODELINGANDANALYSIS

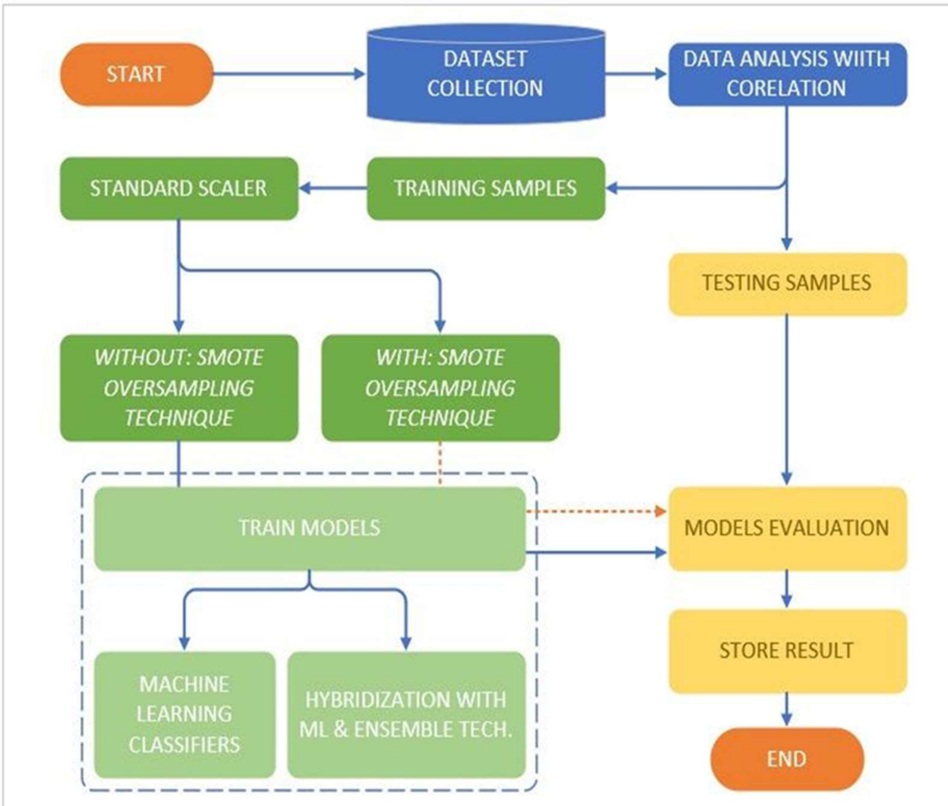


Figure 2: Block Diagram of the Model

SMOTE :Synthetic Minority Oversampling Technique

Oversampling

SMOTE (synthetic minority oversampling technique) is one of the most commonly used oversampling methods to solve the imbalance problem.

It aims to balance class distribution by randomly increasing minority class Examples by replicating them.

SMOTE synthesises new minority instances between existing minority instances. It generates the virtual training records by linear interpolation for the minority class. These synthetic training records are generated by randomly selecting one or more of the k-nearest neighbors for each example in the minority class. After the oversampling process, the data is reconstructed and several classification models can be applied for the processed data.

Random Oversampling includes selecting random examples from the minority class with replacement and supplementing the training data with multiple copies of this instance, hence it is possible that a single instance may be selected multiple times.

Under sampling

Random Undersampling is the opposite to Random Oversampling. This method seeks to randomly select and remove samples from the majority class, consequently reducing the number of examples in the majority class in the transformed data.

ThevariousMachineLearningModelsusedare:

LOGISTICREGRESSION:

Logistic regression is one of the simpler classification models. Because of its parametric nature it can to someextent be interpreted by looking at the parameters making it useful when experimenters want to look atrelationships between variables.The name logistic regression is a bit unfortunate since a regression model isusually used to find a continuous response variable, whereas in classification the response variable is discrete.The term can be motivated by the fact that we in logistic regression found the probability of the responsevariable belonging to a certain class. Thebeta parameter, or coefficient, in this model is commonly estimatedviamaximumlikelihoodestimation(MLE).Oncetheoptimalcoefficient(orcoefficientsif thereismorethanone independent variable) is found, the conditional probabilities for each observation can be calculated, logged,and summed together to yield a predicted probability. For binary classification, a probability less than .5 willpredict 0 while a probability greater than 0 will predict 1. After the model has been computed, it's best practicetoevaluatethehowwellthemodelpredicts thedependentvariable,whichiscalled goodnessoffit.

K-NEARESTNEIGHBOUR:

KNN This section describes the implementation details of KNN algorithm. The model for K-Nearest Neighbor isthe entire training dataset. When a prediction is required for a unseen data instance, the KNN algorithm willsearch through the training dataset for the k-most similar instances. For classification problems, a class label isassigned on the basis of a majority vote-i.e. the label that is most frequently represented around a given datapointisused.Whilethisistechnicallyconsidered“pluralityvoting”,theterm,“majorityvote”ism orecommonly used in literature. The distinction between these terminologies is that “majority voting” technicallyrequires a majority of greater than 50%, which primarily works when there are only two categories. When youhave multiple classes-e.g. four categories, you don't necessarily need 50% of the vote to make aconclusionabouta class;youcouldassignaclasslabel withavote ofgreaterthan25%.

SUPPORTVECTORMACHINE:

SVMaimstofindanoptimalhyperplanethatseparatesthedataintodifferentclasses.Thescikit-learnpackageinpythonisusedforimplementingSVM.Thepre-processeddataissplitintotestdataandtrainingsetwhichisof 25% and 75% of the total dataset respectively. A support vector machine constructs a hyper plane or set ofhyper planes in a high- or infinite-dimensional space. A good separation is achieved by the hyper plane that hasthe largest distance to the nearest training data point of any class (so-called functional margin), since in generalthe larger the margin the lower the generalization error of the classifier. Hyperplanes are decision boundariesthat help classify the data points. Data points falling on either side of the hyperplane can be attributed todifferent classes. Also, the dimension of the hyperplane depends upon the number of features. If the number ofinput features is 2, then the hyperplane is just a line. If the number of input features is 3, then the hyperplanebecomesatwo-dimensionalplane.It becomesdifficultto imaginewhenthenumberof features exceeds.

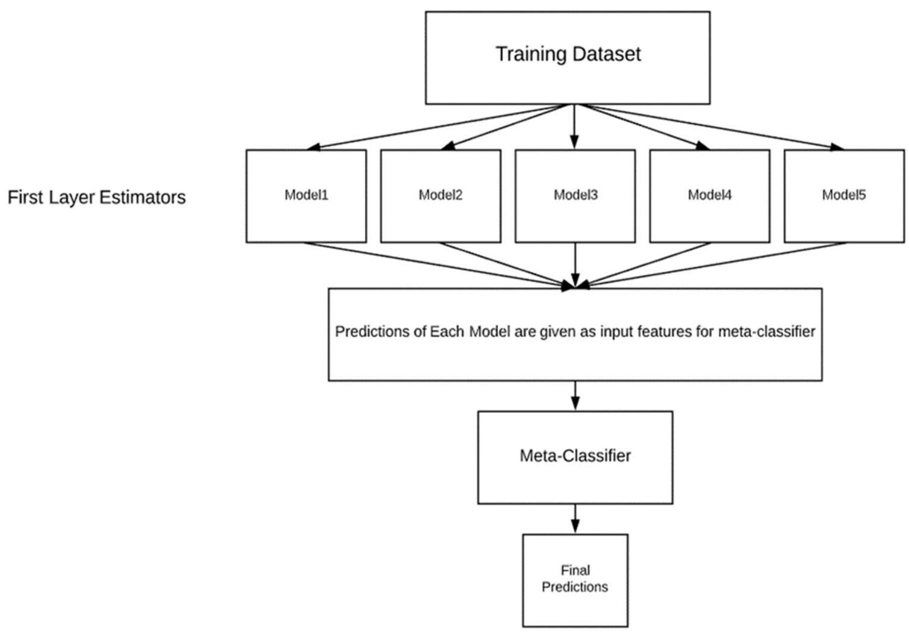
HYBRIDIZATION:

Hybridization is a way of ensembling classification or regression models it consists of two-layer estimators. The first layer consists of all the baseline models that are used to predict the outputs on the test datasets. The second layer consists of Meta-Classifier or Regressor which takes all the predictions of baseline models as an input and generate new predictions. Here I have used three machine learning classifiers like RF,SVC and XGBOOST and make it as hybrid model for liver disease prediction and liver stages prediction.

Specifically, we will evaluate the following 3 algorithms:

- Random Forest
- Support Vector Classifier.
- eXtreme Gradient Boosting Classifier.

Hybride Architecture:



mlxtend:

Mlxtend (machine learning extensions) is a Python library of useful tools for day-to-day data science tasks. It consists of lots of tools that are useful for data science and machine learning tasks for example:

1. Feature Selection
2. Feature Extraction
3. Visualization
4. Ensembling

and many more.

This article explains how to implement Stacking Classifier on the classification dataset.

Why Hybridization ?

Most of the Machine-Learning and Data science competitions are won by using Stacked models. They can improve the existing accuracy that is shown by individual models. We can get most of the Stacked models by choosing diverse algorithms in the first layer of architecture as different algorithms capture different trends in training data by combining both of the models can give better and accurate results.

IV. RESULTSANDDISCUSSION

Our main goal into this Research was to predict liver disease using various machine learningtechniques.We predicted using Hybrid ensemble classifier and it gives 99.96 % of accuracy with better results. I have compare my Proposed Hybrid Classifier with NB, SVM, LOR,RF,DT,KNN, RBTC algorithms. With Each algorithm, we have observed Accuracy, Precision,Sensitivity andSpecificityasfollows:

```
HYBRID CLASSIFIER Accuracy is :99.96%
```

```
from sklearn.metrics import classification_report
STK_Pred=STK.predict(X_test)
STKreport = classification_report(Y_test, STK_Pred)
print(STKreport)
```

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      3561
           1       1.00      1.00      1.00      3489

    accuracy                           1.00      7050
   macro avg       1.00      1.00      1.00      7050
weighted avg       1.00      1.00      1.00      7050
```

Figure 3: Classification Report of Liver Disease Prediction

```
HYBRID CLASSIFIER Accuracy is :98.39%
```

```
from sklearn.metrics import classification_report
STK_Pred=STK.predict(X_test)
STKreport = classification_report(y_test, STK_Pred)
print(STKreport)
```

```
              precision    recall  f1-score   support

         1.0       1.00      1.00      1.00         8
         2.0       0.97      0.97      0.97        62
         3.0       0.98      0.98      0.98       101
         4.0       1.00      1.00      1.00        77

    accuracy                           0.98       248
   macro avg       0.99      0.99      0.99       248
weighted avg       0.98      0.98      0.98       248
```

Figure 4: Classification Report of Liver Disease Stages Prediction

```
confusion matrix for HYBRID CLASSIFIER :
<sklearn.metrics._plot.confusion_matrix.Confusion
```

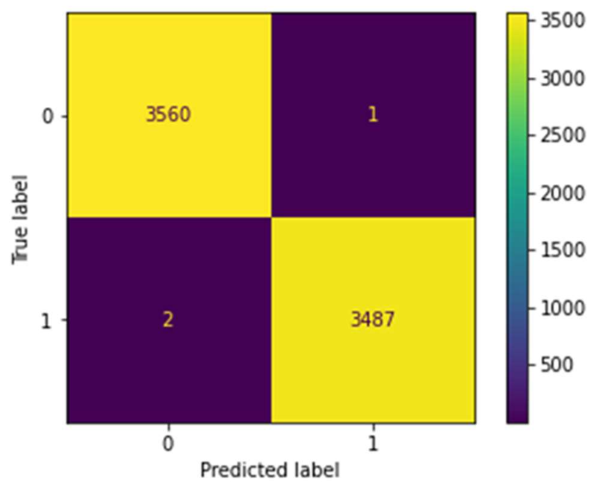


Figure 4: Confusion Matrix of Liver Disease Prediction

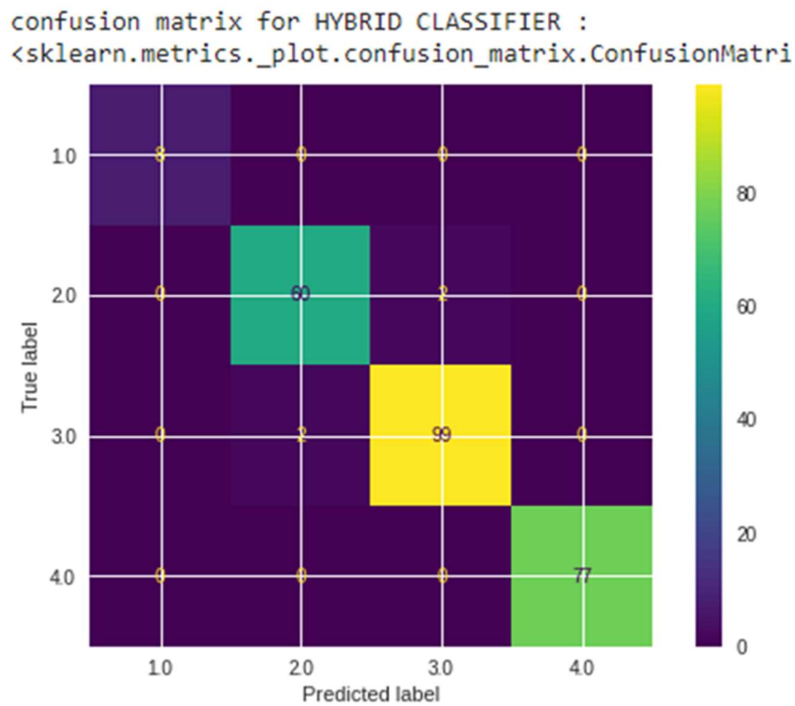


Figure 5: Confusion Matrix of Liver Disease Stages Prediction

V. COMPARISON CHART

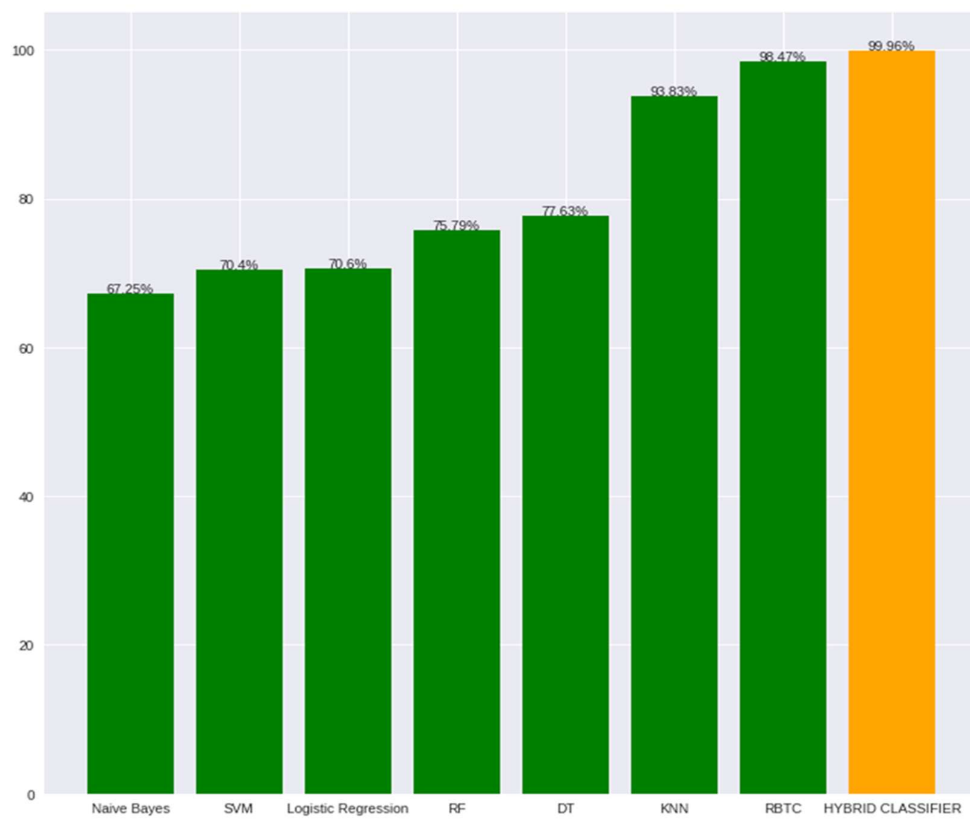


Figure 6: Comparison Chartof Liver Disease Prediction Using Hybrid Classifier.

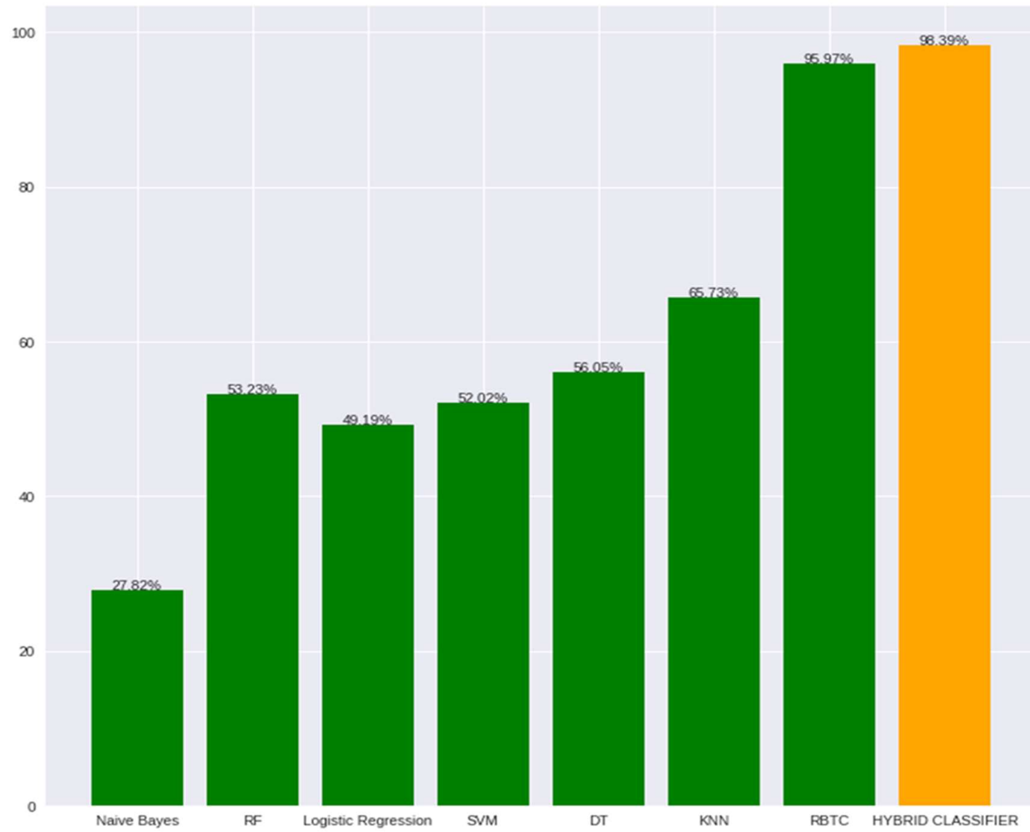


Figure 6: Comparison Chartof Liver Disease Stage Prediction Using Hybrid Classifier.

VI. CONCLUSION

In this research, we have proposed methods for diagnosing liver disease and liver diseases stage prediction in patients using Machine learningtechniques. The many machine learning techniques that were used include SVM,RF, DT,NB , Logistic Regression , KNN,RFBTC and Hybrid Classifier. The system has been implemented using all the models and their performance wasevaluated .The Performance evaluation was based on certain performance metrics. Our Hybridization of RF, SVC and XGBOOST is the proposed model thatresultedinhighestaccuracywithanaccuracyof99% predict the accuracy and give 98% of accuracy to identify a particular stage in liver diseases.

VII. REFERENCES

[1] Yuan-Xing Liua, Xi Liua, Chao Cena, Xin Li b, Ji-Min Liuc, Zhao-Yan Ming d, Song-Feng Yua,Xiao-Feng Tanga, Lin Zhoua, Jun Yua, Ke-Jie Huang b, Shu-Sen Zhenga, “Comparison and development of advanced machine learning tools to predict nonalcoholic fatty liver disease: An extended study” © ELSEVIER 2021.

[2] Jagdeep Singha, SachinBaggab, RanjodhKaurc, “Software Based Prediction of Liver Disease with Feature Selection and Classification Techniques ” © ELSEVIER 2020.

[3] Shivangi Gupta,GreeshmaKaranth,NiharikaPentapati,V R Badri Prasad, “A Web Based Framework for Liver Disease Diagnosis using Combined Machine Learning Models” © IEEE 2020.

[4] Maria Alex Kuzhippallil, Carolyn Joseph,Kannan A, “Comparative Analysis of Machine Learning Techniques for Indian Liver Disease Patients” © IEEE 2020.

[5] PushpendraKumar, Ramjeevan Singh Thakur,“Liver disorder detection using variable-neighbor weighted fuzzy K nearest neighbor approach” 18th International Conference” SPRINGER

[6] PaulR.Harper,Areviewandcomparisonofclassificationalgorithmsformedicaldecisionmaking.

[7] BUPALiverDisorderDataset.UCIrepositorymachinelearningdatabases.

[8] ProfChristopherN.NewAutomaticDiagnosisofLiverStatusUsingBayesianClassification

[9] Schiff'sDiseasesoftheLiver,10thEditionCopyright©2007Ramana,EugeneR.;Sorrell,MichaelMaddrey, WillisC.

[10] P. Sug, On the optimality of the simple Bayesian classifier under zero-one loss, Machine Learning 29 (2–3)(1997)103–130.

[11] 16thEditionHARRISON’SPRINCIPLESofInternalMedicine.

[12] MichaelJSorich.Anintelligentmodelforliverdiseasediagnosis.ArtificialIntelligenceinMedicine2009;47:53—62.