Chronic Liver Diseases and Stage Detection Using Hybrid Machine Learning Model

Mr. Darshan Patel[1], Dr. Dushyantsinh Rathod[2]
PhD Scholar[1] , Professor & HOD[2],
Computer Engineering Department[1,2]
Swarnim Startup & Innvoation Universtiy, Ahmedabad[1],
Ahmedabad Institute of Technology, Ahmedabad, Gujarat[2]
darshanpatel.kse@gmail.com[1], dushyantsinh.rathod@gmail.com[2]

Abstract-Liver disease risk has been rising rapidly among people over the past several decades and is thought to be one of the world's most lethal diseases. To forecast the disease using vast medical datasets is a challenging issue for academics. They have developed machine learning strategies like classification and clustering to address this problem. The primary goal of this research is to use classificational algorithms to predict a patient's likelihood of getting liver disease. Additionally, it indicates the stage of the liver illness, such as Cirrhosis Liver, Liver Fibrosis, Fatty Liver, and Healthy Liver. Accordingly, the suggested Hybrid Classifier (RF,SVC,XGBoost) and the algorithms NB, SVM, LOR,RF,DT,KNN, and RBTC are evaluated for classification accuracy and processing speed. Using these With 99% accuracy, the Hybrid Classifier, which is a superior classifier, is selected after taking performance aspects into account.
Keywords: Logistic Regression, Neural Network, Dataset, Accuracy, SVM, HYBRID model.

I. Introduction

This Research provides the software which facilitates to upload the details and get to know the prediction for Liver disease. This Research uses Machine Learning algorithms to classify whether the liver condition is normal. We use NB, SVM, LOR,RF,DT,KNN, RBTC and Hybridmodels for the prediction. This model will be useful for health industries who need to predict the diseases. The model will be helpful to know whether the liver condition is normal orabnormal using the blood reports of the patient. This information regarding the patients willbe helpful for the medical companies in the process. The existing models include various machine learning techniques which yield output so less accuracy and can't handle large bundles of data.The performance in the training and testing of the liver datasets is observed. These previously designed systems have been sufficient but more work has to be done on their prediction rate for better accuracy in the diagnosis of the liver disease. The proposed system here uses concept of machine learning, and the models are first trained, then tested. Finally the most accurate model will predict the final result. Initially, the system asks you to enter your details including age, gender, total Bilirubin, direct Bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT and Alkphos. These values can be known by blood test report of the user. After taking theseinputs from the user,the system compares the data input with the training dataset of mostaccurate model and then predicts the result 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 Indian Liver Patient Dataset (ILPD) which was selected from UCI Machine learning repository. It is a sample of the entire Indian population collected from Andhra Pradesh region and comprises of 585 patients data. The system is very simple in design and to implement. The system requires very low system resources and the system will work in almost all configurations.

Methodology

The various stages involved are:
Exploratory Data Analysis
Data visualization: With the help of data visualization, we can see how the data looks like and what kind of correlation is held by the

attributes of data. It is the fastest way to see if the featurescorrespond to the output features.

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

Data Preprocessing

This involves eliminating the null and most common words from the text. The words in the dataset consists of links, multiple full stops, very long and short words. These all need to be eliminated before providing it to the

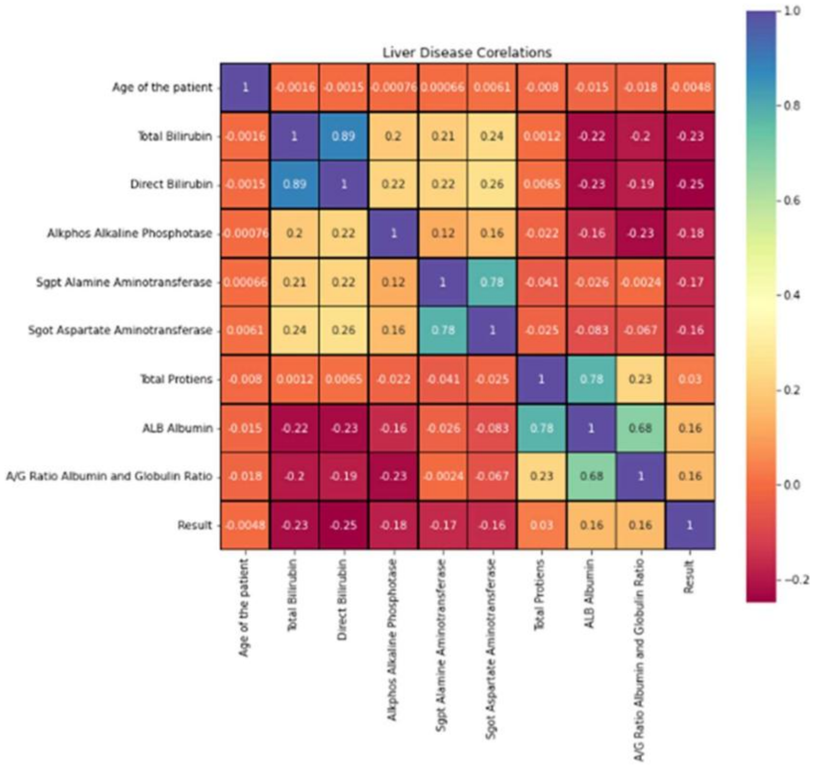

Figure1:Correlation Matrix of the Model

algorithm. The significant stages in data preprocessing are Data Cleaning, Data Integration, Data Reduction and Data Transformation. It is carried out to meet the criteria of accuracy, completeness, consistency, timeliness, believe ability and interpretability.

Training Classification Model

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

II. Proposed Architecture

Machine Learning Models used are:

Logistic Regression:

Logistic regression is one of the simpler classification models. Because of its parametric nature it can to some extent be interpreted by looking at the parameters making it useful when experimenters want to look at relationships between variables.The name logistic regression is a bit unfortunate since a regression model is usually used to find a continuous response variable, whereas in classification the response variable is discrete.The term can be motivated by thefact that we in logistic regression found the probability of the response variable belonging to a certain class. The beta parameter, or coefficient, in this model is commonly estimated via maximum likely hood estimation(MLE).Once the optimal coefficient(or coefficients if there is

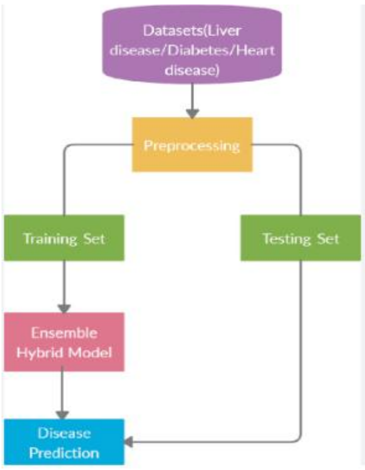

Figure 2: Block Diagram of the Model

more than one 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 will predict 0 while a probability greater than 0 will predict 1. After the model has been computed, it's best practice to evaluate the how well the model predicts the dependent variable, which is called goodness off it.

K-nearest neighbour:

KNN This section describes the implementation details of KNN algorithm. The model for K- Nearest Neighbor is the entire training dataset. When a prediction is required for a unseen data instance, the KNN algorithm will search through the training dataset for the k-most similar instances. For classification problems, a class label is assigned on the basis of a majority vote-i.e. the label that is most frequently represented around a given data point is used.While this is technically considered" plurality voting",the term,"majority vote"is more commonly used in literature. The distinction between these terminologies is that "majority voting" technically requires a majority of greater than 50%, which primarily works when there are only two categories. When you have multiple classes-e.g. four categories, you don't necessarily need 50% of the vote to make a conclusion about a class; you could assign a class label with a vote of greater than25%.

Support vector machine:

SVM aims to find an optimal hyper plane that separates the data in to different classes.The scikit- learn package in python is used for implementing SVM. The pre- processed data is split into test data and training set which is of 25% and 75% of the total dataset respectively. A support vector machine constructs a hyper plane or set of hyper planes in a high- or infinite-dimensional space. A good separation is achieved by the hyper plane that has thelargest distance to the nearest training data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier. Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes. Also, the dimension of the hyperplane depends upon the number of features. If the number of input features is 2, then the hyperplane is just a line. If the number of input features is 3, then the hyper plane becomes a two- dimensional plane. It becomes difficult to imagine when the number of 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:

- RF
- SVM
- XGB.

III. Results

Our main goal into this Research was to predict liver disease using various machine learning techniques. 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 and Specificity as follows:

```
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

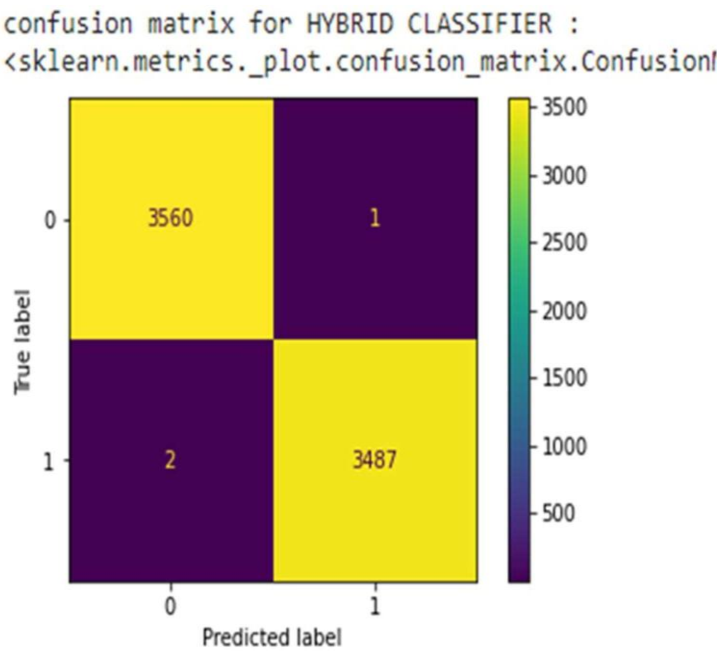

Figure 4: Confusion Matrix of Liver Disease Prediction

IV. Comparison Chart

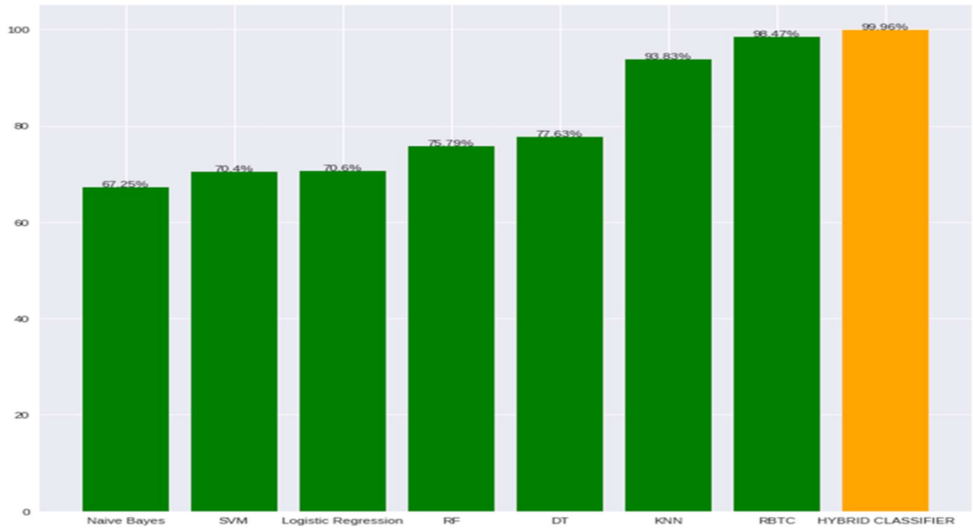

Figure 5: Comparison Chart of Liver Disease Prediction Using Hybrid Classifier

V. Conclusion

In this research, we have proposed methods for diagnosing liver disease and liver diseases stage prediction in patients using Machine learning techniques. 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 was evaluated .The Performance evaluation was based on certain performance metrics. Our Hybridization of RF, SVC and XGBOOST is the proposed model that result gives highest accuracy with an accuracy of 99% predict the accuracy.

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] Paul R.Harper,Are view and comparison of classification algorithms for medical decision making.

[7] BUPA Liver Disorder Dataset. UCI repository machine learning databases.

[8] Prof Christopher N.New Automatic Diagnos is of Liver Status Using Bayesian Classification

[9] Schiff's Diseases of the Liver,10thEdition Copyright©2007 Ramana, EugeneR.;Sorrell,Mi chael Maddrey, 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] 16th Edition HARRISON'SPRINCIPLES of Internal Medicine.

[12] Michael J Sorich.Anintelligent model for liver disease diagnosis. Artificial Intelligence in M edicine2009;47:53—62.