LIVER DISEASE PREDICTION USING ENSEMBLE TECHNIQUES

Dr.C. Prema[1], S. Gayathri[2], P. Sneka[3], B. Snekarani[4], N. Sujitha[5]

[1] Faculty, [2,3,4,5] UG Scholar
Computer Science and Engineering
Jayaraj Annapackiam CSI College of Engineering, Nazareth, India.

Abstract - People have disorder of liver that require medical care at correct time. It is almost important to find the disease before it elapses the curable stage. Significantly, much of understanding of organ development has arisen from analysis of patients with liver deficiencies. Data science is beneficial to find the disease at early stage based on the factors that can be gathered by performing blood test on the patient. Liver disease can be identified by blood test report parameters such as age, gender, total bilirubin, direct bilirubin, alkaline phosphates, alamine aminotransferase, aspartate aminotransferase, total proteins, albumin, albumin and globulin ratio are higher than normal level. In this work, we used three algorithms Logistic Regression, Support Vector Machine, Random Forest. The analysis result shows that the SVM with hyper tuning parameter increase the accuracy. Moreover, our present study mainly focused on the use of clinical data for liver disease prediction and explores different ways of representing such data through our analysis. In This application using various machine learning algorithms find the best accurate to predict on the liver disease or no liver disease.

.

Index Terms - Machine learning, Liver Disease, Logistic Regression, Support Vector Machine, Random Forest, Hyper Tuning Parameter.

I. INTRODUCTION

A. Background Information: To detect disease, healthcare professionals need to collect samples from patients, which can cost both time and money. Often, more than one kind of test or many samples are needed from the patient to accumulate all the necessary information for a better diagnosis. Liver diseases cause millions of deaths every year. Liver failures are at high rate of risk among Indians. It is expected that by 2025 India may become the World Capital for Liver Diseases. India is contributed due to deskbound lifestyle, increased alcohol consumption and smoking. There are about 100 types of liver infections. Using ML, we can predict the liver disease in advance by applying different machine learning algorithm.

B. Machine Learning: Machine learning has become one of the most evolving technologies in the Current period. Machine learning can simply have explained as scientific study of algorithms and models in statistics where machines can easily understand to perform and solve specific tasks. This technique has agile and it has been a requirement in most of the fields.

C. Our Motive: Our motive for this project is to predict the liver disease for a patient with the maximum amount of accuracy in our prediction. For this we have, collected dataset named Indian Patient Liver Disease dataset from UCI repository and used that dataset in our three modules to predict the liver disease using various machine-learning algorithms.

D. Patients Samples: These LFT (Liver Functional Test) include tests like Age, Gender, Total bilirubin, direct bilirubin, alkaline phosphates, Alamine aminotransferase, Aspartate aminotransferase, Total proteins, Albumin, Albumin and Globulin ratio. In this, Software Applications predict the liver disease yes or no. For Examples, Higher levels of AST than ALT can mean alcoholic liver disease. When ALT and AST are increased equally, fatty liver or non-alcoholic liver disease may be the case.

E. Increased Accuracy: ML algorithms are new techniques to handle many hidden problems in medical data sets. This approach can help healthcare management and professionals to explore better results in numerous clinical applications, such as medical image processing, language processing, and tumor or cancer cell detection, by finding appropriate features. Several statistical and machine learning approaches (e.g., simulation modeling, classification, and inference) have been used by researchers and lab technicians for better prediction. The purpose of this study is to propose a new solution for liver disease diagnosis based on SVM, Random Forest, and Logistic Regression. The analysis result shown the SVM achieved the highest accuracy. The major reason to use the SVM is that the result that we are going to obtain through this will be more robust and the results that are obtained through SVM classifiers are accurate that is the reason we have given the SVM priority in our work.

II. LITERATURE SURVEY

In [1] Survival analysis is the most typical methodology in the medication field. Predicting life expectancy is very important factor in both patient and doctor decision-making. In addition to the correct diagnosis of ailment, it is necessary to start appropriate and effective treatment with accurate classification and evaluation it is considered typical for classifying end-stage disease patients. Fuzzy logic is a techniques used to efficiently model inaccurate and complex systems. For the Child-Pugh classification of patients affected by liver cirrhosis of liver, intelligent procedure, supported formal logic could be used. The life expectancy also differs for different classes of patients with cirrhosis patients the life expectancy also differs.

In [2] Early prediction of sickness in incredibly necessary to save lots of human life and take correct steps to regulate the disease call tree algorithms are with success applied in varied fields, particularly in life science, this analysis works explores the first prediction of disease exploitation varied call tree techniques. The disease dataset that is chosen for this study is consisting of attributes like total animal pigment, direct animal pigment, age, gender, and total proteins, simple proteins and simple

proteins magnitude relation. The most purpose of this work is to calculate the performance of varied call tree techniques and compare their performance. The choice tree techniques employed in this study area unit J48, LMT, Random Forest, Random Tree, REPTree, call stump and Hoeffiding Tree. The analysis proves that call Stump Provides the very best accuracy than different techniques.

In [3] Many people are consuming alcohol in the current period, now a day's alcohol consumption is directly associated with threading liver diseases called cirrhosis. Early detection of liver disease is caused by excessive alcohol consumption would help many people save lives It can be diagnosed in time by detecting liver disease at its early stage and can lead to full recovery in some patients. This paper proposes that the presence of liver disease be detected and predicted using data mining algorithms. We will make a dataset decision tree and then generate the rules to coach and check the dataset.

In [4] Data mining has recently improved the simplicity of use for disease prediction in the healthcare sectors. The process of extracting information from huge datasets, warehouses, or other repositories is known as data mining. Predicting diseases using the vast medical datasets is an extremely difficult task for academics. There searchers employ data mining techniques including classification, clustering, association rules, and others to address this problem. This study's primary goal is to use classification algorithms to predict liver disorders. Naive Bayes algorithms were employed in this study. Based on their performance characteristics, such as classification accuracy and execution time, these classifier algorithms are contrasted.

III. METHODOLOGY

In this project, we gather data from a data set, and the health specialist can enter the data for testing using our web application. In this application, we perform data cleaning and pre-processing, extensive data analysis, data visualization, and machine learning using supervised learning algorithms, such as Random Forest, logistic regression, and support vector machines. This approach makes predictions about a person's liver condition based on variables including total bilirubin, direct bilirubin, albumin, total protein, etc.

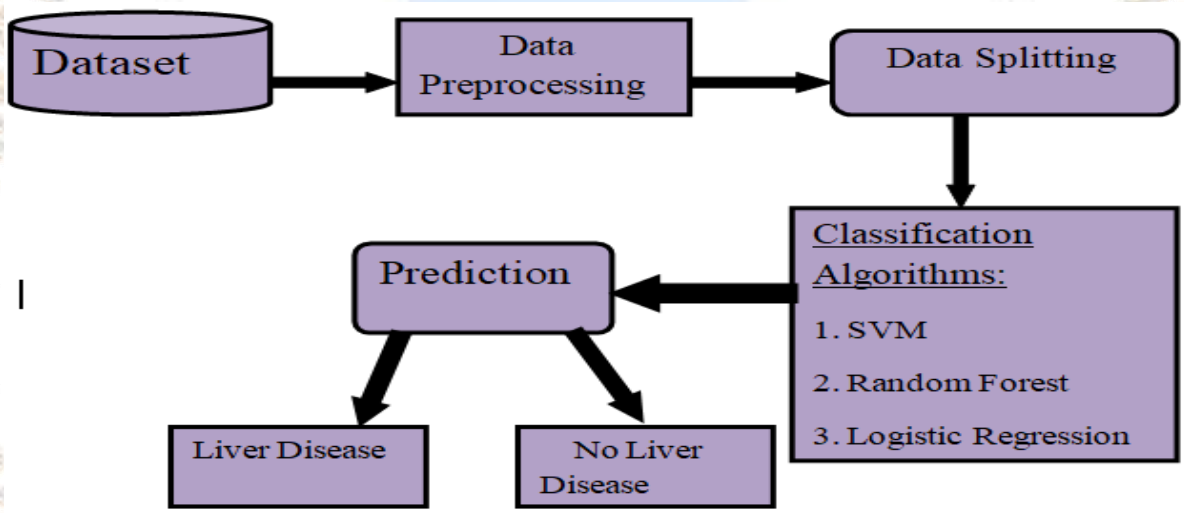

Figure1: Block Diagram

3.1 Dataset collection:

Dataset collection involves the need for selecting appropriate data for analysis and obtaining effective knowledge by performing diverse data mining techniques. The data used for research is Indian Liver Disease Patients (ILDP) from UCI repository.

```
In [5]: #Top 5 rows of the dataset
        liver.head()
```

Out[5]:

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

Figure2: Sample Dataset

3.2 Data Pre-processing:

3.2.1 Imputation of Missing values:

It refers to identifying missing values in the data and imputing the empty values with mean and median values. For Indian Liver Disease Patients data, Albumin and Globulin ratio has four missing values that are replaced by mean values.

```
In [9]: #Checking for null values
        liver.isnull().sum()

Out[9]: 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
```

Figure3: Missing value of Albumin and Globulin ratio

```
In [12]: # Filling NaN Values of "Albumin_and_Globulin_Ratio" feature with Mean :
         liver['Albumin_and_Globulin_Ratio'] =liver['Albumin_and_Globulin_Ratio'].fillna(liver['Albumin_and_Globu
         liver.isnull().sum()

Out[12]: 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
```

Figure4: Replace Missing value of Albumin and Globulin ratio

3.2.2 Label Encoding:

Another data pre-processing technique includes label-encoding data, which focuses on converting the data into machine-readable form. Label encoding converts the labels into numeric forms. In the data used, Gender attribute has labeled data, which is converted in to values 0 and 1(Male 1 and Female 0) for better analysis.

```
In [18]: # Label Encoding
         liver['Gender'] = liver['Gender'].apply(lambda x:1 if x=='Male' else 0)
         liver.head()
```

Out[18]:

	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase
0	65	0	0.7	0.1	187	16	18
1	62	1	10.9	5.5	699	64	100
2	62	1	7.3	4.1	490	60	68
3	58	1	1.0	0.4	182	14	20
4	72	1	3.9	2.0	195	27	59

Figure5: Label Encoding for Male and Female

3.2.3Elimination of duplicate values-

In order to improve the efficiency and quality of data.

3.2.4Outlier detection and Elimination-

Outliers are extreme values that significantly deviate from the rest of the values, which is caused due to inappropriate measurement or experimental error. Different types of outliers include univariate and multivariate outliers. Univariate outliers consider a single feature where as Multivariate outliers look at n-dimensional space consisting of features or attributes of ILPD data. For univariate outlier detection, skewness of attribute is observed and extreme value is replaced. For multivariate outlier detection, isolation forest algorithm is used to identify the contaminated data and it is deleted.

```
In [41]: # removing outliers from certain features and printing len after outliers have been removed
         print('Original dataset:',len(liver))
         liver1 = liver1[liver1.Total_Bilirubin < 40]
         liver1 = liver1[liver1.Direct_Bilirubin < 15.0]
         liver1 = liver1[liver1.Alkaline_Phosphotase < 1250]
         liver1 = liver1[liver1.Alamine_Aminotransferase < 1000]
         liver1 = liver1[liver1.Aspartate_Aminotransferase < 2000]
         liver1 = liver1[liver1.Albumin_and_Globulin_Ratio < 2.0]
         print('After removing outliers:',len(liver1))

         Original dataset: 583
         After removing outliers: 560
```

Figure6: Removing of Outliers.

3.3 Splitting into train and test dataset:

The procedure involves taking a dataset and dividing it into two subsets. The first subset is used to fit the model and is referred to as the training dataset. The second subset is not used to train the model; instead, the input element of the dataset is

provided to the model, then predictions are made and compared to the expected values. This second dataset is referred to as the test dataset.

- ❖ Train Dataset: Used to fit the machine learning model.
- ❖ Test Dataset: Used to evaluate the fit machine learning mode.

Figure7: Splitting the Training and Testing

3.3 Classification algorithm:

SVM algorithm*:*

SVM is a supervised machine learning algorithm, which can be used for classification or regression problems. It uses a technique called the kernel trick to transform our data and then based on these transformations it finds an optimal boundary between the possible outputs. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyper plane.

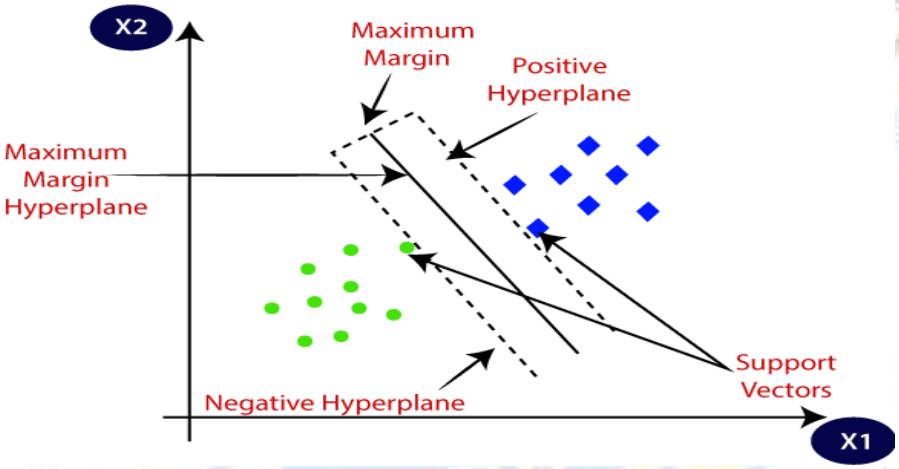

Figure8: Block Diagram of SVM

Logistic Regression:

Logistic regression is a supervised classification algorithm used to predict the probability of a target variable. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, and cancer detection. Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

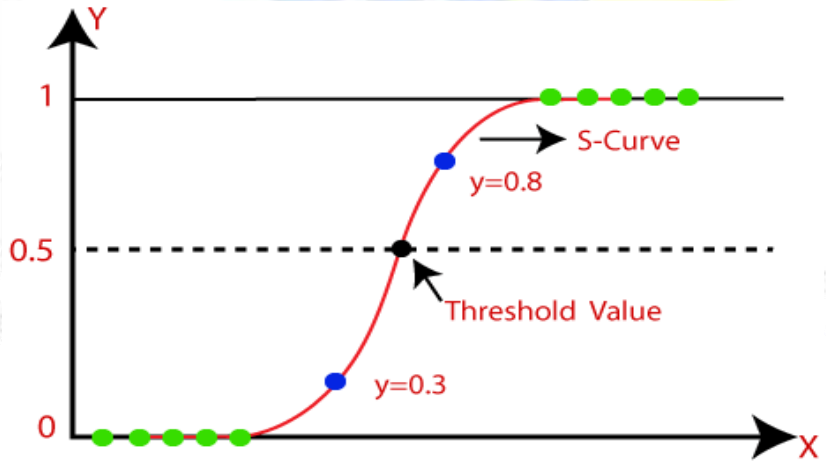

Figure9: Block Diagram of Logistic Regression

- ➢ **Logistic Function:**
  - The sigmoid function is a mathematical function used to map the predicted values to probabilities.
  - It maps any real value into another value within a range of 0 and 1.

Random Forest Algorithm:

Random forests are one of the ensemble learning methods for classification (and regression) that works by generating multiple decision trees at training time and showing outcome the class result by individual trees. It is an excellent algorithm in terms of accuracy among mentioned algorithms. It runs effectively on large database. It can handle thousands of input

variables without variable shrinking. It gives estimates of variables having importance in the classification. Random Forests generated number of classification trees. The forests then have a choice of the classification having the most votes.

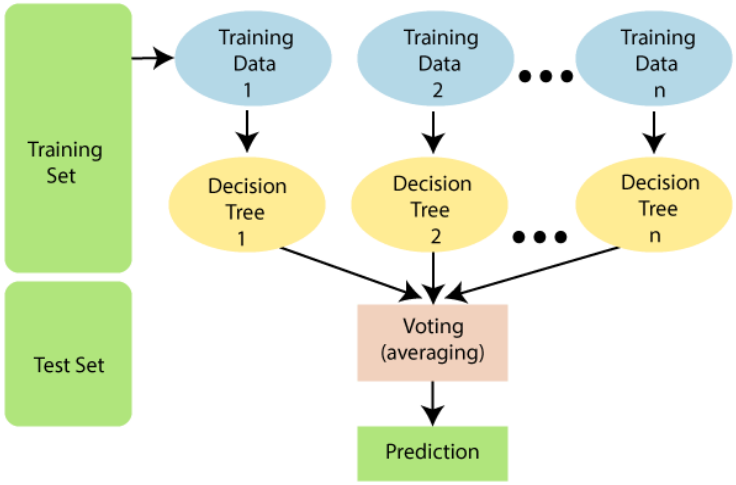

Figure10: Block Diagram of Random Forest

IV.RESULTS AND DISCUSSION

Our goal is to get prediction on the basis of given datasets of people whether the person is having the liver disease or no liver disease. With using many different algorithms, such as SVM Algorithm using hyper tuning parameter, Logistic Regression and Random Forest we are trying to predict which algorithm will play vital role in predicting the disease.

SVM Algorithm:

```
[ ]  # accuracy score
     svc_score=round(svc.score(X_train,y_train)*100,2)
     print(svc_score)
     svc_acc = round(accuracy_score(y_test,y_pred)*100,2)
     print(svc_acc)

     69.1
     80.34

     # classification report
     print(classification_report(y_test, y_pred))

                   precision    recall  f1-score   support

                0       0.00      0.00      0.00        23
                1       0.80      1.00      0.89        94

         accuracy                           0.80       117
        macro avg       0.40      0.50      0.45       117
     weighted avg       0.65      0.80      0.72       117

[ ]  # confusion matrix
     print(confusion_matrix(y_test, y_pred))
     sns.heatmap(confusion_matrix(y_test,y_pred),annot=True,fmt="d")

     [[ 0 23]
      [ 0 94]]
```

Figure11: Result SVM using hyper tuning Algorithm

Random Forest Algorithm:

```
[ ]  # accuracy score
     rand_clf_score=round(rand_clf.score(X_train,y_train)*100,2)
     print(rand_clf_score)
     ran_clf_acc = round(accuracy_score(y_test, y_pred)*100,2)
     print(ran_clf_acc)

     95.49
     76.07

[ ]  # classification report
     print(classification_report(y_test, y_pred))

                   precision    recall  f1-score   support

                0       0.35      0.26      0.30        23
                1       0.83      0.88      0.86        94

         accuracy                           0.76       117
        macro avg       0.59      0.57      0.58       117
     weighted avg       0.74      0.76      0.75       117

     # confusion matrix
     print(confusion_matrix(y_test, y_pred))
     sns.heatmap(confusion_matrix(y_test,y_pred),annot=True,)

     [[ 6 17]
      [11 83]]
     <matplotlib.axes._subplots.AxesSubplot at 0x7f5a179f10a0>
```

Figure12: result on Random Forest Algorithm

Logistic Regression:

```
[ ]  # accuracy score
     lr_score=round(lr.score(X_train,y_train)*100,2)
     print(lr_score)
     lr_acc = round(accuracy_score(y_test,y_pred )*100,2)
     print(lr_acc)

     72.32
     76.92

     # classification report
     print(classification_report(y_test, y_pred))

                   precision    recall  f1-score   support

                0       0.36      0.22      0.27        23
                1       0.83      0.90      0.86        94

         accuracy                           0.77       117
        macro avg       0.59      0.56      0.57       117
     weighted avg       0.73      0.77      0.75       117

[ ]  # confusion matrix
     print(confusion_matrix(y_test, y_pred))
     sns.heatmap(confusion_matrix(y_test,y_pred),annot=True,fmt="d")

     [[ 5 18]
      [ 9 85]]
     <matplotlib.axes._subplots.AxesSubplot at 0x7f5a17bb9c10>
```

Figure13: Results on Logistic Regression

This system aims to predict the person liver condition with higher accuracy so we have compared three algorithms out of which we have chosen SVM algorithm as the best one as it gives the higher accuracy. The various results of the algorithms are shown as below.

SL.NO	Classification Algorithm	Accuracy
1.	SVM	80.34%
2.	Logistic Regression	76.92%
3.	Random Forest	76.07%

Table1: Results of classification algorithms

4.1Sample Screenshots:

Here the value 1 means that the prediction is “Liver Disease”

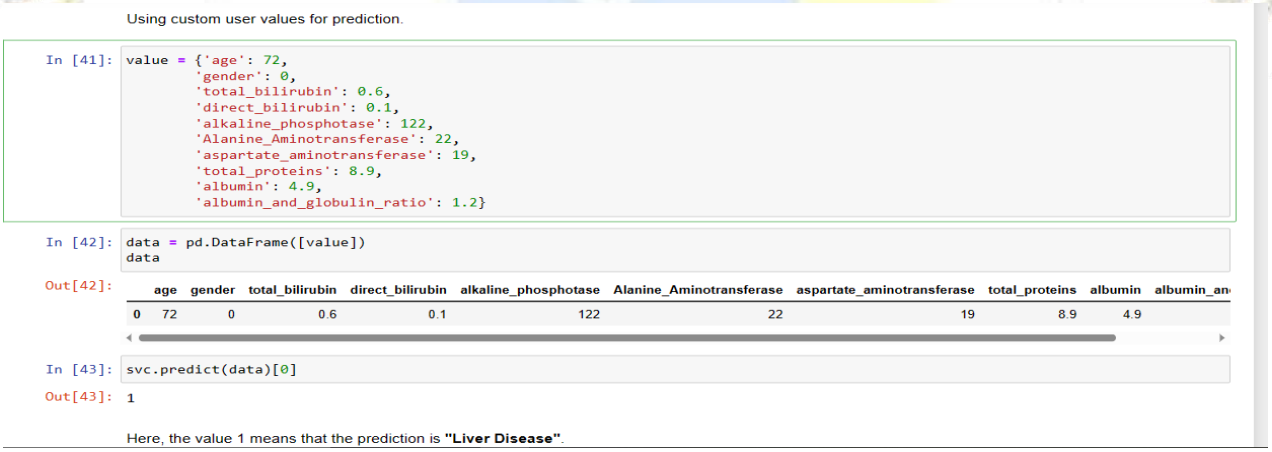

Figure 14: prediction of Liver Disease

Here the value 0 means that the prediction is “No Liver Disease”

```
In [91]: value = {
             "Age": 17,
             "Gender": 1,
             "Total_Bilirubin": 0.9,
             "Direct_Bilirubin": 0.3,
             "Alkaline_Phosphotase": 202,
             "Alamine_Aminotransferase": 22,
             "Aspartate_Aminotransferase": 19,
             "Total_Protiens": 7.4,
             "Albumin":4.1,
             "Albumin_and_Globulin_Ratio":1.2
         }
         data = pd.DataFrame([value])
         data
```

Out[91]:

	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumi
0	17	1	0.9	0.3	202	22	19	7.4	4.1	

```
In [92]: print(rand_clf.predict(data)[0])
```

```
0
```

Figure 15: Prediction of no liver disease

V.FUTURE ENCHANCEMENT

This work presents an approach that will be used for hybrid model construction of community health services. These classification algorithms can be implemented for other dominant diseases also like cardiac and diabetes prediction and classification. More than one dataset may be used for better approach and comparison. Another scope is to see whether by applying new algorithms will result any improvements over techniques, which are used in this work in future. More techniques for accuracy increment may be applied. Wrapper method may be applied for removing noise in the dataset.

Classification rules and disease identifying techniques may also be generated by using different efficient algorithms. More than one database for comparative analysis may also be used. Our works has certain limitations as the model has underperformed having less accuracy than expectations. Therefore, in future, inclusion of deep learning methods may improve our results further.

VI. REFERENCES

[1]. Fuzzy Logic for Child –Pugh classification of patients with Cirrhosis of Liver, Anu Sebastian, Surekha Mariam Varghese. [IEE-2016]
[2] Liver Disease Prediction by using different Decision Tree Techniques, Nazmun Nahar and Ferdous Ara [Research gate, 2018]
[3] Liver Disease Detection Due to Excessive Alcoholism Using Data Mining Techniques, Insha Arshad, Chiranjit Dutta. [IEEE, 2018]
[4] Liver Disease Prediction using Naïve Bayes Algorithms, Dr. S. Vijayarani 2000.
[5] Gulia A, Vohra R, and Rani P. (2014). Liver Patient Classification Using Intelligent Techniques. International J. of Computer Science and Information Techno-logies, 5(4), 5110-5115.
[6] Mostafa, F, Hasan E, Williamson M,Khan H, Statistical Machine Learning Approaches to Liver Disease Prediction. Livers **2021**, 1, 294–312. https://doi.org/10.3390/livers1040023
[7] Siva Kumar D, Manjunath Varchagall, and Ambika L Gusha S "Chronic Liver Disease Prediction Analysis Based on the Impact of Life Quality Attributes." (2019). International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7, Issue-6S5, April 2019.
[8]Rahman, A. S., Shamrat, F. J. M., Tasnim, Z., Roy, J. and Hossain, S. A. (2019). A Comparative Study on Liver Disease Prediction Using Supervised Machine Learning Algorithms. International J. of Scientific &Technology Research, 8(11), pp.419-422.