PERFORMANCE ANALYSIS OF DIFFERENT MACHINE LEARNING ALGORITHM FOR PREDICTION OF LIVER ANOMALIES

A. Joel Dickson [a], J. Sophia Jone [b], R. Renuka [c], S. Kipsy [d]

[a] *Assistant professor, Department of ECE, Bethlahem Institute of Engineering, Kanyakumari, Tamil Nadu, India*
[b] *Assistant professor, Department of ECE, Bethlahem Institute of Engineering, Kanyakumari, Tamil Nadu, India*
[c] *PG Students – Department of ECE, Bethlahem Institute of Engineering, Kanyakumari, Tamil Nadu, India*
[d] *PG Students – Department of ECE, Bethlahem Institute of Engineering, Kanyakumari, Tamil Nadu, India*

ABSTRACT: ***The aim of my project is to provide the maximum accuracy of the prediction of liver diseases using various machine learning techniques. Data Mining is one in every of the foremost vital aspects of automatic malady identification and malady prediction. It involves data processing algorithms and techniques to research medical knowledge. In the recent times, the percentage of liver disease patients have been numerously increasing and this liver related diseases have been a fatal diseases now-a-days. In this project various machine learning techniques are used to predict the liver disease for a person. In this project, I have collected a liver disease patient dataset from the kaggle. This project is implemented in three various modules. In the first module, we have to clean the dataset. Then min-max algorithm to the dataset. In the next phase, classification of the dataset is done with various machine learning algorithms. In the last phase various machine learning techniques to predict liver disease.***

I. INRODUCTION

A. Machine Learning

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

B. Aim

The aim of this project is to predict the liver diseases for a patient with the maximum amount of accuracy in our prediction. Liver is the largest internal organ in the human body, it is essential for digesting food and releasing the toxic element of the body and plays a major role in metabolism and serving several vital functions. The liver is the largest glandular organ of the body. It weighs about 3 lb. (1.36 kg). The liver's main job is to strain the blood coming from the digestive tract, before passing it to the rest of the body. The liver also detoxifies chemicals and metabolizes drugs. As it does so, the liver hides bile that ends up back in the intestines. The liver also makes proteins important for blood clotting and other functions. The liver supports almost every organ in the body and is vital for our survival. Liver disease may not cause any symptoms at earlier stage or the symptoms may be vague, like weakness and loss of energy. Symptoms partly depend on the type and the extent of liver disease. Liver diseases are diagnosed based on the liver functional test. Several diseases states can disturb the liver. Some of the diseases are Wilson's disease, hepatitis (an inflammation of the liver), liver cancer, and cirrhosis (a chronic inflammation that progresses ultimately to organ failure). Alcohol alters the metabolism of the liver, which can have on the whole detrimental effects if alcohol is taken over long periods of time. Hemochromatosis can cause liver problems

Patient samples: Since the number of percentage of people with liver disease have been increasing and there are very less efficient methods to predict liver diseases. The methods we use in this project would bring more efficiency the prediction methods. To do so we have to collect the different patients' samples and make it as a dataset. Then we need to refine the samples of the patients we have collected to increase the accuracy percentage.

Efficient Technique: Here in our project, we have identified a best algorithm which can give a better accuracy compared to other machine learning

techniques using the liver dataset. The algorithm we have identified is gradient booster.

Increased Accuracy: We had many algorithms in the past for detecting the disease but the algorithms which we have used in our project will increase the efficiency in predicting the liver disease. The machine learning algorithm called gradient booster technique gives us higher accuracy compared to others.

C. Data Mining

Data extraction is the way to find designs in expansive informant indexes including AI crossing point strategies, measurements and database frames. Information Mining is an interdisciplinary field of software engineering and measuring that aims to separate data from information collection (with keen strategies) and transform data into a understandable structure to be used further.

D. Data Pre-processing

Data pre-processing is an important step to solve each problem of machine learning. In order for a machine learning algorithm to be trained, most of the data sets used with machine teaching problems must be processed. The techniques used most commonly for pre- processing are very few such as imputation of lack of value, categorical coding, scaling, etc. These are easy to understand techniques. When we deal with the data, however, things often become cumbersome.

Every dataset has its own unique challenges and is different. All features, except Gender are real valued integers. The last column, Disease, is the label (with '1' representing presence of disease and '2' representing absence of disease). Total number of data points is 583, with 416 liver patient records and 167 non-liver patient records

Dataset: The Indian Liver Patient Dataset contained 10 distinct qualities of 583 patients. The patients were portrayed as either 1 or 2 based on liver sickness. The nitty gritty portrayal of the dataset is appeared Table. The table give insights regarding the trait and characteristic sort. As plainly unmistakable from the table, every one of the highlights with the exception of sex are genuine esteemed numbers. The component Sex is changed over to numeric esteem (0 and 1) in the information pre-preparing step.

Data Collection: Collection of data is crucial for these kinds of projects. We have collected a dataset named as Indian Patient Liver Dataset from UCI repository which consists of 10 different attributes of 583 patients.

Data Cleaning: We will have different columns which are called attributes. Some columns have null values and some values are fluctuated so we will clean those values from the datasheet and then take that dataset for classification.

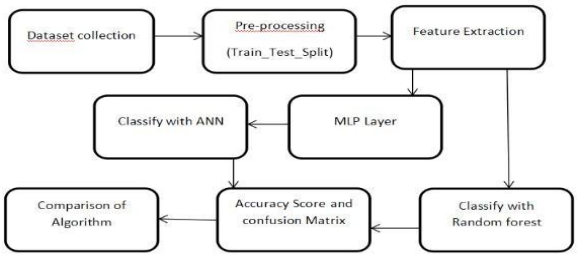
II. SYSTEM ANALYSIS

A. Proposed System

PSO feature extraction model for the liver dataset is applied to improve the chances in abundant medical applications like coaching artificial neural networks, linear unnatural operational improvement, wireless network improvement, information classification, and lots of different areas wherever GA will be applied. Computation in PSO relies on a swarm of process parts known as particles within which every particle represents a candidate answer.

A multilayer perceptron (MLP) is a neural network model which can map liver datasets of input file onto a collection of applicable outputs. Associate MLP classification is multiple nodes in an exceedingly directed plot, with every layer absolutely connected to successive one. In this proposed system, we have to collect a liver disease patient dataset from the UCI repository.

This project is implemented in three various modules. In the first module, we have to clean the dataset which we have collected from the UCI repository. Then we have to apply the min- max algorithm to the dataset. In the next phase, we will do the classification of the dataset which we have collected from the UCI repository. The next phase is the phase three where we will have to apply various machine learning techniques to predict liver disease.



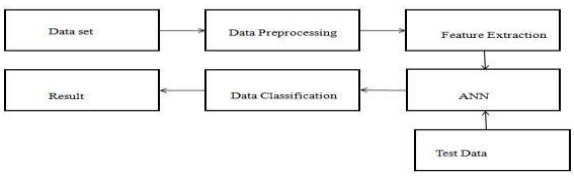
Flow chart of proposed model

B. Issues in existing methodology

The existing methodology had an algorithm called SVM algorithm to calculate that disease details. SVM algorithm is slow algorithm for classifying and it also gives less accuracy. In that main disadvantage is time efficiency. Details of patient liver diseases start from large scale, diverse, fully independent and distributed and seek to explore complex and evolving patterns between data. Existing system shows an algorithm called SVM [Support Vector Machine] theorem that characterizes the options of the massive information revolt, and implement an enormous processing model.

C. New Methodology

PSO feature extraction model for the liver dataset is applied to improve the chances in abundant medical applications like coaching artificial neural networks, linear unnatural operational improvement, wireless network improvement, information classification, and lots of different areas wherever GA will be applied. Computation in PSO relies on a swarm of process parts known as particles within which every particle represents a candidate answer. A multilayer perceptron (MLP) is a neural network model which can map liver datasets of input file onto a collection of applicable outputs. Associate MLP classification is various deep layers of nodes in an exceedingly directed plot, with every layer absolutely connected to successive one.



System Architecture

III. SYSTEM DESIGN

A. Algorithm

SVM Algorithm: SVM algorithm tries to give out hyper planes and split the data into different categories. The scikit-learn package in python is employed for implementing SVM. The pre-processed information is split into check information and coaching set that is of twenty fifth and seventy fifth of the entire dataset severally. A SVM technique builds hyper planes in an exceedingly dimensional area. a decent separation is achieved by the hyper plane that has the most important distance to the closest coaching information of any category (so-called purposeful margin), since generally the larger the margin the lower the generalization error of the classifier.

KNN: KNN is algorithm which is also called as the K-nearest neighboring algorithm. It is also called as the lazy learning algorithm. Its motivation is to utilize a database in which the information focuses are isolated into a few classes to anticipate the order of another example point.

Random Forest: Random Forest is also called as Random Decision Trees. Random Forest algorithm is a machine learning technique where these are a learning tasks, classification and regression tasks.

Decision Trees: A decision tree is a flowchart-like structure in which each interior hub speaks to a "test" on a characteristic (for example, regardless of whether a coin flip comes up heads or tails), each branch speaks of the result of the test, and each leaf node speaks to a class mark (choice taken in the wake of processing all qualities). The ways from root to leaf speak to characterization rules.

Gradient Booster: Gradient boosting is an AI system for relapse and order issues, which delivers an expectation show as a group of feeble forecast models, commonly decision trees. It manufactures the model in a phase savvy style like other boosting strategies do, and it sums them up by permitting enhancement of a self-assertive differentiable lost function.

B. Modules & Functionalities

As mentioned, there are three modules in our project. These modules are divided into three phases such as data preprocessing, classification and the other module is algorithm implementation.

C. Dataset

The Indian Liver Patient Dataset contained 10 distinct qualities of 583 patients. The patients were portrayed as either 1 or 2 based on liver sickness. The nitty gritty portrayal of the dataset is appeared Table. The table give insights regarding the trait and characteristic sort. As plainly unmistakable from the table, every one of the highlights with the exception of sex are genuine esteemed numbers. The component Sex is changed over to numeric esteem (0 and 1) in the information pre-preparing step.

Data Pre-processing: Data pre-processing is an important step of solving every machine learning problem. Most of the datasets used with Machine Learning problems need to be processed / cleaned / transformed so that a Machine Learning algorithm can be trained on it. Most commonly used pre-processing techniques are very few like missing value imputation, encoding categorical variables, scaling, etc. These techniques are easy to understand. But when we actually deal with the data, things often get clunky. Every dataset is different and poses unique challenges. All features, except Gender are real valued integers. The last column, Disease, is the label (with ’1’ representing presence of disease and ’2’ representing absence of disease). Total number of data points is 583, with 416 liver patient records and 167 non-liver patient records. In the description of this dataset, it is observed that some values are Null for the Albumin and Globulin Ratio column. The columns which contain null values are replaced with mean values of the column.

Classification: In this data classification module we have a trained dataset and a test dataset. Firstly, we will classify the dataset into train and test datasets. Then we have to train the dataset and we next take train dataset at 80% and test dataset at 20% and then we will apply algorithms to it.

Algorithm implementation: In this module we have implemented different algorithms with the dataset. The first step is we have trained the dataset. In this implementation part we have 80% of the train dataset and 20% of the test dataset. We have implemented various algorithms like random forest, decision trees, gradient booster, KNN algorithm, naive bayes.

IV. DATA PRE-PROCESSING

```
In [32]: df=pd.read_csv('indian_liver_disease.csv')

In [33]: df.shape
Out[33]: (583, 11)

In [34]: df.columns
Out[34]: Index(['Age', 'Gender', 'Total_Bilirubin', 'Direct_Bilirubin',
                'Alkaline_Phosphotase', 'Alamine_Aminotransferase',
                'Aspartate_Aminotransferase', 'Total_Proteins', 'Albumin',
                'Albumin_and_Globulin_Ratio', 'Dataset'],
               dtype='object')

In [35]: df.head()
Out[35]:
```

	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Proteins	Albumin	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	

Attributes of the dataset

A. Exploratory Data Analysis

Filtering categorical data

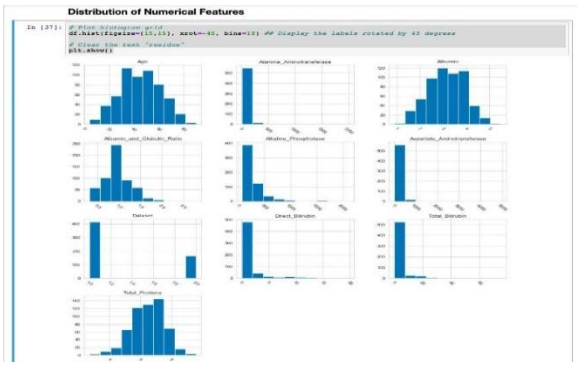
df.dtypes [df.dtypes=='object']

Exploratory Data Analysis

```
Filtering categorical data

In [36]: df.dtypes[df.dtypes=='object']
Out[36]: Gender    object
         dtype: object
```

B. Distribution of Numerical Features

df.hist(figsize=(15,15), xrot=-45, bins=10) ## Display the labels rotated by 45 degress plt.show()



Distribution Of Numerical Features

C. Mean values

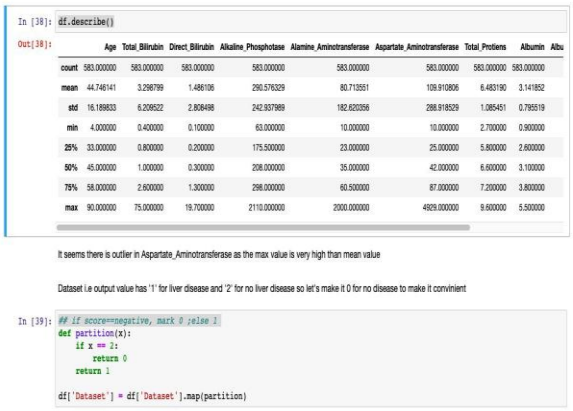
For displaying the mean and standard deviation. And it seems there is outlier in Aspartate Aminotransferase as the max value is very high than mean value Dataset i.e output value has '1' for liver disease and '2' for no liver disease so let's make it 0 for no disease to make it convenient def partition(x):

if x == 2:

return 0

return 1

df['Dataset'] = df['Dataset'].map(partition)



```
In [38]: df.describe()
```

Out[38]:

	Age	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Proteins	Albumin	Albu
count	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000	
mean	44.746141	3.298799	1.486106	290.576329	80.713551	109.910806	6.483190	3.141852	
std	16.189833	6.209522	2.808498	242.937989	182.620356	288.918529	1.085451	0.795519	
min	4.000000	0.400000	0.100000	63.000000	10.000000	10.000000	2.700000	0.900000	
25%	33.000000	0.800000	0.200000	175.500000	23.000000	25.000000	5.800000	2.600000	
50%	45.000000	1.000000	0.300000	208.000000	35.000000	42.000000	6.600000	3.100000	
75%	58.000000	2.600000	1.300000	298.000000	60.500000	87.000000	7.200000	3.800000	
max	90.000000	75.000000	19.700000	2110.000000	2000.000000	4929.000000	9.600000	5.500000	

It seems there is outlier in Aspartate_Aminotransferase as the max value is very high than mean value

Dataset i.e output value has '1' for liver disease and '2' for no liver disease so let's make it 0 for no disease to make it convinient

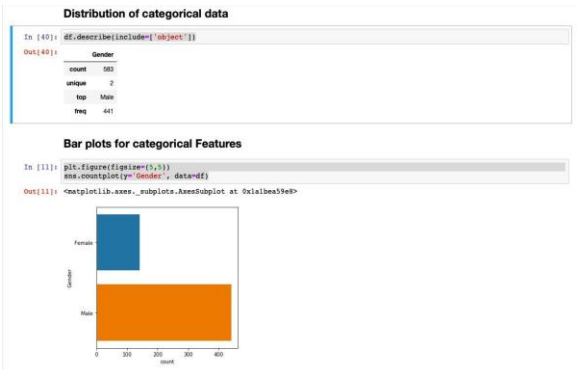
```
In [39]: ## if score==negative, mark 0 ;else 1
def partition(x):
    if x == 2:
        return 0
    return 1

df['Dataset'] = df['Dataset'].map(partition)
```

D. Distribution of categorical data

df.describe(include=['object'])

Command:
plt.figure(figsize=(5,5))

sns.countplot(y='Gender', data=df)**E. Machine**



Distribution of categorical data

```
In [40]: df.describe(include=['object'])
```

Out[40]:

	Gender
count	583
unique	2
top	Male
freq	441

Bar plots for categorical Features

```
In [11]: plt.figure(figsize=(5,5))
         sns.countplot(y='Gender', data=df)
Out[11]: <matplotlib.axes._subplots.AxesSubplot at 0x1a1bea59e8>
```


E. Learning Models

Data Preparation

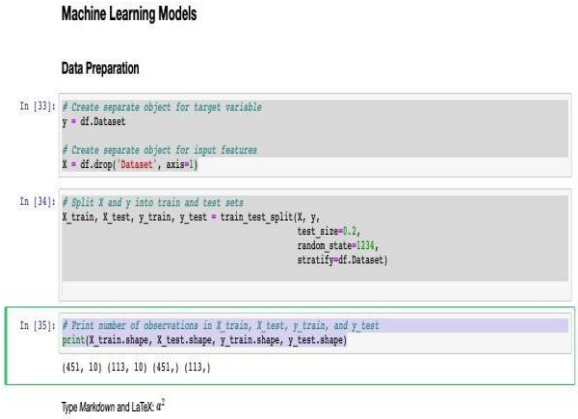
Command

Print (X_train.shape, X_test.shape, y_train.shape, y_test.B shape)

Description: To Create separate object for target variable And to Create separate object for input features Split X and y into train and test sets

And to Print number of observations in X_train, X_test, y_train, and y_test

Screenshot:



Machine Learning Models

Data Preparation

```
In [33]: # Create separate object for target variable
         y = df.Dataset

         # Create separate object for input features
         X = df.drop('Dataset', axis=1)
```

```
In [34]: # Split X and y into train and test sets
         X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                             test_size=0.2,
                                                             random_state=1234,
                                                             stratify=df.Dataset)
```

```
In [35]: # Print number of observations in X_train, X_test, y_train, and y_test
         print(X_train.shape, X_test.shape, y_train.shape, y_test.shape)

(451, 10) (113, 10) (451,) (113,)
```

Type Markdown and LaTeX: α^2

F. Data standardization

In Data Standardization we perform zero mean centering and unit scaling; i.e., we make the mean of all the features as zero and the standard deviation as 1. Thus we use mean and standard deviation of each feature. It is very important to save the mean and standard deviation for each of the feature from the training set, because the same mean & std deviation in the test set.

Command

```
train_mean = X_train.mean()

train_std = X_train.std()

X_train = (X_train - train_mean) / train_std

X_train.describe()

X_test = (X_test - train_mean) / train_std

X_test.describe()
```

G. Testing

Black-box Testing: Testing is a part of every project to find out bugs. All the projects need testing but types of testing for different projects is different. It is also called transparent testing. It is used for the internal structures of the application and also working of an application. It is also a functional testing. It also involves functional part of the application

TESTED ID	ALGORITHM	ACTUAL OUTPUT	PREDICTED OUTPUT	SUCESS RATE
1	Neutral network	100	71.52	SUCCESS
2	Logistic regression	100	71.0	SUCCESS
3	Decision tree	100	66.2	SUCCESS
4	Random forest	100	73.0	SUCCESS
5	Gradian boost	100	76.8	SUCCESS

V. CONCLUSION

Through this project we have increased the efficiency of the prediction. We have increased the accuracy of the prediction algorithms where we have used different algorithms to predict the accuracy of the disease at different accuracy levels. We have used a specific dataset Indian liver patient dataset where we have 10 attributes and more than 500 patients' data so it would be very useful and give best accuracy of the prediction.

REFERENCES

1. Fuzzy Logic for Child-Pugh classification of patients with Cirrhosis of Liver AUTHORS: Anu Sebastian, Surekha Mariam Varghese
2. Liver disease prediction by using different decision tree techniques
3. Liver Disease Detection Due to Excessive Alcoholism Using Data Mining Techniques AUTHORS: Insha Arshad, Chiranjit Dutta.
4. Nonalcoholic Fatty liver disease. American Family Physician. 2013;88(1).
5. Lactulose for hepatic encephalopathy. Medical Letter on Drugs and Therapeutics.
6. Propranolol doses not decrease the development of large esophageal varices in patients with cirrhosis A controlled study. Hepatology. 1995; 22.
7. First Definition of Reference Intervals of Liver Function Tests in China: A Large-Population-Based Multi-Center Study about Healthy Adults
8. M. Abdel-Basset, et al., 2-Levels of clustering strategy to detect and locate copy-move forgery in digital images. Multimedia Tools and Applications, 1–19, 2018.
9. M. Abdel-Basset, et al., Internet of Things (IoT) and its impact on supply chain: A framework for building smart, secure and efficient systems, Future Generation Computer Systems, 2018.
10. Abdar, M. et al., Performance analysis of classification algorithms on early detection of liver disease. Expert Syst. Appl. 67:239–251,2017.
11. Muktevi srivenkatesh, et al., Performance evolution of different machine learning algorithms for prediction of liver disease, 2019
12. Shah nazir, et al., performance assessment of classification algorithms on easy detection of liver syndrome. 2020
13. Elias Dritsas, et al., Supervised machine learning methods for liver disease risk perdition. 2023
14. Nahian, et al., Common human disease prediction using machine learning based on survey data. 2022
15. Fahad Mostafa, statistical machine learning approaches to liver disease prediction. 2021