

**PERFORMANCE ASSESSMENT OF VARIOUS MACHINE LEARNING  
METHODS FOR PREDICTION LIVER DISEASE**

**BY**

**Arafatur Rahman Soikat  
ID: 221-25-111**

This Report Presented in Partial Fulfillment of the Requirements for the  
Degree of Master of Science in Computer Science and Engineering

Supervised By

**Shah Md Tanvir Siddiquee**  
**Assistant Professor**  
Department of CSE  
Daffodil International University

Co-Supervised By

**Mr. Narayan Ranjan Chakraborty**  
**Associate Professor**  
Department of CSE  
Daffodil International University



**DAFFODIL INTERNATIONAL UNIVERSITY**

**DHAKA, BANGLADESH**

**JANUARY 2023**

## APPROVAL

This Thesis titled “**Performance Assessment of Various Machine Learning Methods for Prediction Liver Disease**”, submitted by **Arafatur Rahman Soikat**, ID No: **221-25-111** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 17-01-2023.

### BOARD OF EXAMINERS



**Dr. Touhid Bhuiyan, PhD**  
**Professor and Head**

Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University

Chairman



**Ms. Nazmun Nessa Moon**  
**Associate Professor**

Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University

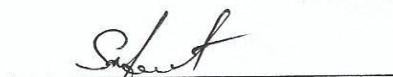
Internal Examiner



**Dr. Fizar Ahmed**  
**Associate Professor**

Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University

Internal Examiner



**Md. Safaet Hossain**  
**Associate Professor & Head**

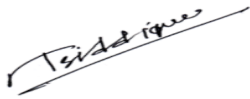
Department of Computer Science and Engineering  
City University

External Examiner

## DECLARATION

I hereby declare that this project has been done by us under the supervision of **Shah Md Tanvir Siddiquee, Assistant Professor**, and Department of CSE Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of this degree.

**Supervised by:**



**Shah Md Tanvir Siddiquee**  
**Assistant Professor**  
Department of CSE  
Daffodil International University

**Co-Supervised by:**



**Mr. Narayan Ranjan Chakraborty**  
**Associate Professor**  
Department of CSE  
Daffodil International University

**Submitted by:**



**Arafatur Rahman Soikat**  
ID: 221-25-111  
Department of CSE  
Daffodil International University

## ACKNOWLEDGEMENT

First, I express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

I really grateful and wish our profound our indebtedness to of **Shah Md Tanvir Siddiquee, Assistant Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “Machine Learning” to carry out this research. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this research.

I would like to express our heartiest gratitude to **Dr. Touhid Bhuiyan, Professor and Head**, Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

I would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, I must acknowledge with due respect the constant support and patients of our parents.

## **ABSTRACT**

Chronic Liver Disease (CLD) is the leading cause of death worldwide, affecting a large number of people. A variety of factors damage the liver, resulting in this disease. Obesity, undiagnosed hepatitis, and alcohol abuse are only a few examples. This is the cause of irregular nerve activity, blood in the cough or vomit, kidney failure, liver failure, jaundice, liver encephalopathy, and many other symptoms. Therefore, the goal of this work is to evaluate the best algorithm find from different types of prediction algorithm using the values of predicted accuracy. In this work, I used six algorithms Logistic Regression, K Nearest Neighbors, Decision Tree, Support Vector Machine, Naive Bayes, and Random Forest. Different measurement techniques, such as accuracy, precision, recall, f-1 ranking, and specificity, were used to assess the performance of different classification techniques. The performance parameters, such as classification accuracy and execution time, are used to compare these classifier algorithms. According to the findings of the experiments, the RF and KNN are best classifier algorithms for predicting liver diseases.

## TABLE OF CONTENTS

<b>CONTENTS</b>	<b>PAGE</b>
Board of examiners	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
<b>LIST OF FIGURES</b>	viii
<b>LIST OF TABLES</b>	ix
<b>CHAPTER</b>	
<b>CHAPTER 1: INTRODUCTION</b>	<b>1-5</b>
1.1 Introduction	1
1.2 Motivation	3
1.3 Rational of the study	4
1.4 Research Questions.	5
1.5 Expected Output	5
<b>CHAPTER 2: BACKGROUND</b>	<b>6-10</b>
2.1 Introduction	6
2.2 Related Work	6
2.3 Comparative Analysis and Summary	9

2.4	Scope of the Problem	10
2.5	Challenges	10
<b>CHAPTER 3: RESEARCH METHODOLOGY</b>		<b>12-26</b>
3.1	Introduction	12
3.2	Research Design	13
3.3	Data Collection Procedure	14
3.4	Proposed Methodology	14
3.4.1	Data Mining Tool	15
3.4.2	Data Processing	15
3.5	System Design	18
3.6	KNN Algorithm	20
3.7	Logistic Regression	20
3.8	Decision Tree	21
3.9	SVM Algorithm	21
3.10	Naïve Bayes	22
3.11	Random Forest Algorithm	22
3.12	Logical Data Model	23
3.13	Algorithm Decision Structure	24
3.14	Pair Plot Graph between Each Entities	25
<b>CHAPTER 4: IMPLEMENTATION AND RESULT ANALYSIS</b>		<b>27-41</b>

4.1	Introduction	27
4.2	Six Machine Learning Prediction Algorithm Applying procedure	27
4.3	Experimental Result and Analysis	28
4.3.1	Naïve Bayes	28
4.3.2	Logistic Regression	29
4.3.3	Decision Tree	31
4.3.4	SVM	32
4.3.5	KNN	34
4.3.6	Random Forest	36
4.4	Result Analysis	38
4.5	Description of My Work	39
	<b>CHAPTER 5: CONCLUSION AND FUTURE WORK</b>	<b>42-46</b>
5.1	Conclusion	42
5.2	Future Work	45
	<b>REFERENCES</b>	<b>47-49</b>



## LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.1: Research Design	13
Figure 3.2: Here, Total Number of the male and female of this data set	16
Figure 3.3: Bar chart of Total Number of the male and female of this data set	16
Figure 3.4: Total Number of the Ratio of Patient's	17
Figure 3.5: Bar chart of Total Number of the Ratio of Patient's	17
Figure 3.6: System Design	19
Figure 3.7: KNN Algorithm	20
Figure 3.8: Decision Tree	21
Figure 3.9: Logic Data Model	23
Figure 3.10: Pair plot graph between each Entities	25
Figure 3.11: Data Set index	26
Figure 4.1: Visualization between actual value and predicted value for Naïve Bayes	29
Figure 4.2: Visualization between actual value and predicted value for Logistic Regression	30
Figure 4.3: Visualization between actual value and predicted value for Decision Tree	32
Figure 4.4: Visualization between actual value and predicted value for SVM	34
Figure 4.5: Visualization between actual value and predicted value for KNN	36
Figure 4.6: Visualization between actual value and predicted value for 6 Random Forest	38
Figure 4.7: Accuracy Measure Visualization	39

## LIST OF TABLES

<b>TABLES</b>	<b>PAGE NO</b>
Table 1.1: Number of the missing value of each attributes of this data batch.	11
Table 4.1: Confusion Matrix for Naïve Bayes	28
Table 4.2: Classification Report for Naïve Bayes	28
Table 4.3: Confusion Matrix for Logistic Regression	29
Table 4.4: Classification Report for Logistic Regression	30
Table 4.5: Confusion Matrix for Decision Tree	31
Table 4.6: Classification Report for Decision Tree	31
Table 4.7: Confusion Matrix for SVM	33
Table 4.8: Classification Report for SVM	33
Table 4.9: Classification Report for KNN	35
Table 4.10: Confusion Matrix for KNN	35
Table 4.11: Classification Report for Random Forest	37
Table 4.12: Confusion Matrix for Random Forest	37
Table 4.13: Accuracy Measure	38

# CHAPTER 1

## Introduction

### 1.1 Introduction

Liver disease is defined as any interruption in liver function that results in sickness. The liver is in charge of several critical processes in the body, and if it gets diseased or damaged, those activities may be endangered, potentially leading to catastrophic health repercussions. A hepatic problem is a disorder that affects the liver. The phrase "liver disease" refers to a collection of conditions that cause the liver to malfunction. Before the function is compromised, more than 75% of the liver tissue, or three-quarters, must be damaged.

The liver is the most important organ in your body. It assists in food digestion, resource storage, and the removal of pollutants from your body. There are several types of liver problems. Virus-caused liver diseases, such as Hepatitis A, B, and C. Drugs, toxins, or excessive alcohol consumption can all lead to liver damage. Cirrhosis and fatty liver disease are two examples. Hemochromatosis and Wilson disease are inherited disorders. A liver ailment is sometimes known as liver cancer.

Essential liver malignant growth is the fifth most normal disease on the planet and the third primary driver of malignant growth mortality. As per the latest public vault figures in Japan, essential liver malignant growth is the third driving reason for death in men and the fifth driving reason for death in ladies. Hepatocellular carcinomas represent around 90% of all essential liver malignant growths (HCC). In terms of underlying liver disease, the Liver Cancer Research Group of Japan's most recent national study discovered that hepatitis C virus (HCV)-related liver disease is the most prevalent underlying cause of HCC. HCC caused by the hepatitis C virus (HCV) accounts for 67 percent of all HCC, followed by HCC caused by the hepatitis B virus (HBV) at 15%. One out of every ten Americans (30 million people) has liver disease. Cirrhosis, often known as chronic liver disease, affects around 5.5 million Americans.

Several kinds of liver disease are becoming increasingly widespread in the United States as obesity rates rise. Non-alcohol-related fatty liver disease affects 20% to 30% of the population (NAFLD). Because of its connection to metabolic syndrome and illnesses such as diabetes, high blood pressure, high cholesterol, and obesity, this ailment may be referred to as metabolic-associated fatty liver disease (MAFLD).

Some kinds of liver illness, such as non-alcoholic fatty liver disease, have little symptoms. Jaundice, which is characterized by a yellowing of your skin and eyes, is the most common symptom of various disorders.

The frequency of liver disease is rising. Since 1970, deaths from liver disease have increased 400%. Every day, over 40 people in the United Kingdom die as a result of liver illness. In contrast, the number of deaths from other major killer illnesses, such as heart disease and cancer, has either remained constant or decreased. Every year, around 7700 individuals die as a result of alcohol-related liver illness. Hepatitis B is present in 180,000 people, and the hepatitis C virus is present in 143,000 people. It's estimated that 40-50 percent of people with viral hepatitis go undiagnosed. Hepatitis affects 90 percent of people. Every year, around 6000 people are diagnosed with primary liver cancer, or approximately 16 people each day. The great majority of these people would have progressive liver disease as their underlying ailment. Only 12% of persons diagnosed with primary liver cancer live for more than five years. Liver disease deaths in Bangladesh reached 23,143 in 2018, accounting for 2.98 percent of all deaths, according to the most recent WHO reports. Bangladesh is ranked 78 in the world with a death rate of 19.21 per 100,000 people.

If I wish to run liver disease prediction analysis in various patients in the future, I need to know the patient's age, gender, and ethnicity. Total Bilirubin, Direct Bilirubin, Alkphos Alkaline Phosphatase, Sgpt Alamine Aminotransferase, Sgot Aspartate Aminotransferase, Total Protiens, ALB Albumin, A/G Albumin, and Globulin Ratio, Result. A prediction is a number, a series of numbers, or any other related information that is produced in accordance with future events. In comparison to forecasts, the foundation of forecasting is

historical knowledge, which is associated with intuition and skepticism. Here I want to find out this type of number with different patients symptoms.

If I take the weather on a particular patient as an example, I can classify prediction datasets as "Age of the patient", "Gender of the patient", "Total Bilirubin", "Direct Bilirubin", "Alkaline Phosphatase", "SGPT Alanine Aminotransferase", "SGOT Aspartate Aminotransferase", "Total Proteins", "ALB Albumin", "A/G Ratio Albumin" and "Globulin Ratio", "Result". The process of training and mapping datasets in order to obtain detailed study knowledge. Here I used 7 types of algorithm from machine learning for detecting liver disease. The algorithms are Naive Bayes, LR (Logistic Regression), SVM (Support Vector Machine), IBK (K nearest Neighbor), J48 (Decision Tree) and Random Forest.

## **1.2 Motivation**

Bangladesh is home to an expected 8 million individuals who are distressed with persistent hepatitis B and C infections (HCV). Furthermore, common severe hepatitis flare-ups caused by hepatitis A infection (HAV) and hepatitis E infection are unavoidable (HEV). Non-communicable disease patients, such as nonalcoholic fatty liver diseases (NAFLDs), are also on the rise. During the inaugural yearly conference of the South Asian Relationship for the Investigation of the Liver (SAASL) in Dhaka, Bangladesh in 2013, data on the recurrence of liver illnesses in Bangladeshi hepatology departments was gathered.

The Rajshahi division had 19,200 patients and the Rangpur division had 525. They ranged in age from 15 to 95 years old. Individuals under the age of 15 were not allowed to participate. The bulk of the patients were men (67.9 percent). Many of the patients went to their respective hepatology departments expecting to be diagnosed with liver disease. However, further investigation revealed that 13.2 percent of the patients had liver disorders. Medical examination, biochemistry, and imaging were used to make the diagnosis.

The majority of patients with liver disease had chronic liver diseases (CLDs), accounting for 37 to 69 percent of all cases. Complications of CLD, such as hepatic encephalopathy (HE), are less common (2.6%) in locations like Dhaka division, where the healthcare

system is comparable to better, than in places like Khulna division, where the healthcare system is less established. where the healthcare system is somewhat underdeveloped [13]. Many people go to their liver treatment late without realizing it, which causes a lot of financial and physical damage to them. For this reason, they can use it to prevent this problem. They find out the best accuracy of the liver so they can look for medical help before they have more liver problems.

### **1.3 Rational of the study**

Chronic liver disease is a worldwide health burden due to its development from benign and controllable failure to life-threatening end-stage liver disease with catastrophic implications. Transplantation is the only and last solution for end-stage liver failure due to a lack of sufficient therapy. More therapeutic approaches are badly needed to minimize progression and increase end-stage liver disease when demand for organ transplantation considerably exceeds availability. I do this research so that no patient is at this stage. This allows them to understand the current state of their liver in the event of a minor problem and to take precautionary measures before a major problem arises.

Different people have different benefit issues for which I have researched about 30791 patient data. Our research is mainly based on data mining and machine learning. Here, I have researched six algorithms of machine learning (Naive Bayes, LR (Logistic Regression), SVM (Support Vector Machine), KNN (K nearest Neighbor), Decision Tree and Random Forest. In this research I have collected data of different patients, where I have worked with the patient's age, "Age", "Gender", "Total Bilirubin", "Direct Bilirubin", "Alkphos Alkaline Phosphotase", "Sgpt Alamine Aminotransferase", "Sgot Aspartate Aminotransferase", "Total Protiens", "ALB Albumin" and Result (It has two category "1" represent for patient already effected by liver disease and '2' represent for patient aren't effected ) some more necessary data. After all these struggles I find out the best accuracy in KNN, SVM, Decision Tree and random forest algorithm. Those who want to make the best medical instrument on liver disease can get a good output by working on this research.

### **1.4 Research Questions**

In the research, I am attempting to address a variety of questions.

- Why KNN, Decision Tree, SVM and Random Forest best?
- What does KNN, Decision Tree, SVM and Random Forest do?
- What are the advantages of KNN, Decision Tree, SVM and Random Forest?
- How conscious I need to be for liver disease?

### **1.5 Expected Output**

- Finding the actual accuracy similar between real data and test data of liver disease.
- The best algorithm to prediction liver disease.
- If I find the perfect algorithm then the prediction accuracy of liver disease will be close to hundred presents accurate.
- Comparing the results of different prediction algorithm using liver disease dataset from liver patients.
- Execution time analysis and accuracy measurement for liver disease dataset.
- Determine the accuracy of various data mining strategies for liver illness.

## CHAPTER 2

### Background

#### 2.1 Introduction

Best Liver Disease Prediction Algorithm for actual accuracy similar between real data and test data of liver disease. To begin making a prediction, choose some data based on characteristics such as data collecting, data processing, data selection, feature selection, preprocessing, and performance accuracy. Different Decision Frameworks will be described in this part for various applications that are linked to this study. This chapter contains detailed work, linked work, and a summary of the research. Specifics on the scope of the problem. Here are the goals I set for ourselves and the obstacles I faced. A population-based observational cohort have a look at of people in Tayside, Scotland, who had performed in number one care and have been accompanied for two years. To determine the baseline functions of the derivation cohort, biochemistry information were connected to secondary care, prescriptions, and mortality records. To externally examine the very last model, an extraordinary validation cohort turned into obtained from 19 preferred practices across the relaxation of Scotland.

#### 2.2 Related work

In recent years, liver disease prediction researchers have built many systems to determine of finding liver disease prediction. A. K. M Sazzadur Rahman et.al [1] has anticipated six essential illnesses, including distorted nerve capability, hacking up or spewing blood, kidney disappointment, liver disappointment, jaundice, and hepatic encephalopathy, to bring down the significant expense of persistent liver infection determination. They utilized six calculations for sickness expectation: including Nave Bays(53), Logistic Regression (75), Decision Tree (69), Support Vector Machine (64), k Nearest Neighbors (62), and Random Forest (74). A correlation of these calculations was done in view of their characterization exactness measure. As per the preliminary outcomes, calculated relapse procured the best exactness as the predominant calculation that anticipated disease with the



precision of the most noteworthy arrangement than different calculations. Dr. S. Vijayarani et.al [2] has used data mining for liver disease prediction. Data mining has been used to drag up information. The researchers used extensive medical datasets to anticipate the condition. Make use of data mining techniques. To predict liver disease, they employed classification algorithms. These are naïve bays and support vector machines. They have determined that the support vector machine is a better classifier for predict the liver illness. Shapla Rani Ghosh et al. [3] projected a liver cancer disease to medical the accuracy, precision, sensitivity, and specificity of many approaches on hepatic sickness diagnosis such as nave bays classification, bagging, kSatr, logistic, and REP tree on two data sets from UCLA and AP were examined. It was determined that kstar performed with the best accuracy and precision. Based on the trial findings, they discover an increase in the accuracy rate of medical tools to lesion and cost on hepatic illness diagnosis. They achieved the goal by using the kstar algorithm, which can be utilized on diagnostic tools to quickly identify particular liver disorders. Kumar, Yugal ET. AI [4] proposed a rule-based classification model with machine learning technique for the prediction of different types of liver disease and to simplify the basis of result and analysis of laboratory test, different liver disease are classified by using support vector machine, rules induction, Decision Tree, Naive Bays, and Artificial Neural Network with k-cross fold techniques. For predicting liver sickness, the proposed model was used and compared to a model without rules, with the rules-based classification model employing the decision tree technique yielding the most accurate results. Esraa M. Hashem and other AI [5] used a categorization system for cancer medical diagnosis tools. They used diabetic illness datasets and hepatic patent datasets to develop support vector machine for identifying liver disease. Analyze the support vector machine algorithm's performance in terms of accuracy, error rate, sensitivity, prevalence, and specificity. According to the experimental results, first 8order characteristics had the greatest accuracy, error rate, and specificity for diabetes diagnostic datasets. In comparison to the other two datasets. Shahadat Uddin et al. [6] used many supervised machine learning algorithms to forecast a single illness. They analyzed 48 distributions altogether for the pressure among variants regulated AI calculation for

expectation calculation and found that the help vector machine technique (in 29 explorations) and guileless sound calculation are the most ordinarily utilized (in 23 examinations). For the trial results, Irregular Timberland (53) had the most exactness in 9 of the 17 examinations utilized, while the help vector machine had the most noteworthy precision in 41% of the investigations dissected. To improve and guarantee the exactness of liver expectations, a preparation license technique is fundamental. A decision-making method with many criteria that can minimize the amount of certification changes. The approach evaluates internal and external characteristics and prioritizes the coefficients required to make an educated choice using the instrument. It will also assist the authority in developing unique training programs focused on identifying the importance of important decision factors. This approach use a normalized procedure to examine separate measures and combine them to obtain the final forecast. It produces multiple algorithm outcomes for the same data. As a consequence, utilizing a multiplicative method with a classification function overcomes the problems of predicting liver function using an additive approach. The investigation effort investigates the early prediction of liver problems utilizing multiple decision tree Ways. The main goal of this study is to compute and compare the performance of several decision tree methods. They discovered that Decision Stump outperforms other strategies in terms of accuracy [16]. They compare with 12 distinct characteristics. The comparison of two decision tree algorithms, FT heightening and Nave Bayes, and setting up Nave Bayes is better than the FT heightening algorithm with the use of machine learning because Nave Bayes 75.54% delivers greater accuracy than the FT heightening algorithm 72.66% using the WEKA instrument. In contrast, in the balance of the FT tree, Nave Bayes to forecast the liver illness confused with assessment using 10-fold cross-validation [17]. They enhanced the diagnosis of liver illnesses by examining two recognition formulas: patient variable and genome utterance. They also examine and list the drawbacks of the computational algorithms that may be employed in the aforementioned technique. It proposes strategies for increasing the efficiency of these algorithms [18]. They employ an obliquely attached Dense DNN on the 13 specified LFT indicators and individuals' arithmetic knowledge for liver disease screening. The Dense

Deep Neural Network's Area under the Curve is 0.8919, the Deep Neural Network's (0.8867%), the Random Forest's (0.8790%), and the logistic regression's (0.7974%). The performance of deep learning models outperforms that of the traditional formula. In terms of the deep learning formula, the Dense Deep neural network outperforms the deep neural network [19]. They enhanced a large public unit's nomogram-based NAFLD prediction type. We concentrated on surveying machine learning devices for predicting NAFLD [20].

### **2.3 Comparative Analysis and Summary**

Exact liver prediction is required to lower the chance of complete mortality in our human existence, which includes physical, mental, and financial consequences. The liver can be predicted using a liver prediction model. To offer liver numbers, machine learning approaches may be utilized to interpret and assess liver designs. One of the most extensively utilized approaches for liver predicting is information mining. Data mining is a technique for analyzing data and inferring rules that may be used to anticipate events. The main outcome is the best algorithm find from different types of prediction algorithm using the values of predicted accuracy. This research employs a process. I choose criteria, and each criterion has its own set of sub-criteria. Data collection, preparation, visualization, interpretation, and categorization may all be computed [7]. The liver is the human frame's 2d largest internal organ. It is necessary for the human body to supply protein, coagulate blood, and to metabolize cholesterol, glucose, and iron. The liver also has the role of casting off pollutants from the body, making it important for survival. When the liver fails to paintings well, the various body's features are impaired, ensuing in extreme damage to the frame. The liver can be harmed if it becomes infected with a deadly disease, is assaulted through its own immune gadget, or is uncovered to chemical compounds. Hepatotrophic viruses, together with hepatitis B virus (HBV), hepatitis C virus, and hepatitis delta virus, can motive lethal liver harm and cause continual liver disease.

## 2.4 Scope of the Problem

- The amount of data was limited since there was less previous documentation or study articles.
  - As my study data was derived from the Liver Patient Dataset train, various difficulties were encountered in gathering data from there.
    - Because this data collection contains 30691 people's liver information,
    - It was incredibly tough to introduce this strategy.
    - All data is not current or fixed. Because they alter on a daily basis, determining accuracy is tough.
      - This study's raw data processing and equation computations required a long time.
      - Changes in a person's surface highlights affect a variety of variables. For example, they can have an impact on a patient's liver condition. Massive networks can also make it challenging for meteorologists to anticipate narrow-scope hepatic events effectively.
        - For instance, they can influence the condition of liver's patient. They can also make it difficult for liver specialists to accurately predict restricted scope liver events.
          - Changes in a human's surface highlights can have a range of implications. They can also make it harder for liver experts to forecast restricted scope liver events properly.

## 2.5 Challenges

- I need to contact several departments, such as Liver Patient Dataset to gather data, and I also need to graph what I want to achieve in our study and the significance of our work. Effective management of availability, usability, and dramatic review of significant knowledge are the most significant challenges.
- This dataset was compiled from several Liver Patient Dataset sources. Our data collection has 30691 people's liver information, thus collecting data from there has been difficult.
- But there are lots of data was missing values. Total missing values is 5425 which is effected on final accuracy. But I fill this missing data using

`fillna(datasetname[attributename].mean())` technique applying pandas library in python  
 Show a table which attributes data was missing in blow.

TABLE 1.1: NUMBER OF THE MISSING VALUE OF EACH ATTRIBUTES OF THIS  
 DATABATCH

Name of the attributes	Number of the missing value
Age of the patient	2
Gender of the patient	902
Total Bilirubin	648
Alkphos Alkaline Phosphotase	796
Sgpt Alamine Aminotransferase	538
Sgot Aspartate Aminotransferase	462
Total Protiens	463
ALB Albumin	494
A/G Ratio Albumin and Globulin Ratio	559

- Many liver data are secured in kaggle.
- Merge three data set.
- The collection of attributes is a challenging job.
- Each decision is given weight.
- All data are not up to date or corrected. These fluctuate from day to day.
- Data synchronization took time to design as well.
- It is tough to select a decision framework in this system.
- There aren't many works with in-depth information investigation, particularly in the liver figure.
- Finally, I said that best algorithm finds from different types of prediction algorithm using the values of predicted accuracy.

## **CHAPTER 3**

### **Research Methodology**

#### **3.1 Introduction**

Algorithms abound in computer science. For predicting liver disease, I used a machine learning prediction system. This study explores the idea and predicts the accuracy of liver disease such "Correct Classified Instances", "Incorrectly Classified Instances", "Specificity", "Precision", "Sensitivity", "F1 Measure". Logistic regression is used in various fields, including machine learning, most medical fields, and social sciences. Application of Naive Bayes Algorithm real class prediction, multi class prediction, Text classification/ Spam Filtering/ Sentiment Analysis Recommendation System. SVM, or Support Vector Machine, is a linear model that can be used to remedy class and regression problems. It can take care of linear and nonlinear issues and is useful for a wide range of realistic programs. The idea of SVM is straightforward: The method draws a line or a hyperplane to divide the facts into training. The k-nearest naiver algorithm (k-NN) is a non-parametric classification approach invented in 1951 by using Evelyn Fix and Joseph Hodges [1] and later modified by using Thomas Cover. Its packages consist of type and regression. The input in both instances includes the k nearest naiver examples in the statistics set. Decision Trees are a non-parametric supervised learning approach that may be used for classification as well as regression applications. The objective is to build a model that predicts the value of a target variable using basic decision rules derived from data attributes. His random forest is a classification system made up of several decision trees. When creating each individual tree, it employs bagging and feature randomization in an attempt to produce an uncorrelated forest of trees whose forecast by committee is more accurate than that of any individual tree. The liver is in charge of various key tasks in the body, and when it becomes ill or wounded, those functions might be lost, resulting in considerable bodily harm. The liver is the biggest organ in the human body, and its prognosis is critical; nonetheless, there has been little study in this area. I'm undertaking research to see if I can create the best

algorithm for predicting liver illness. Some of my work has already been the subject of some interesting research.

### 3.2 Research Design

Because the goal of this study is to forecast liver illness, the approach used is quantitative, with maximum data collecting from several sources. This dataset was compiled using several sources from the Liver Patient Dataset. Our data collection contains information about 30691 people's livers.

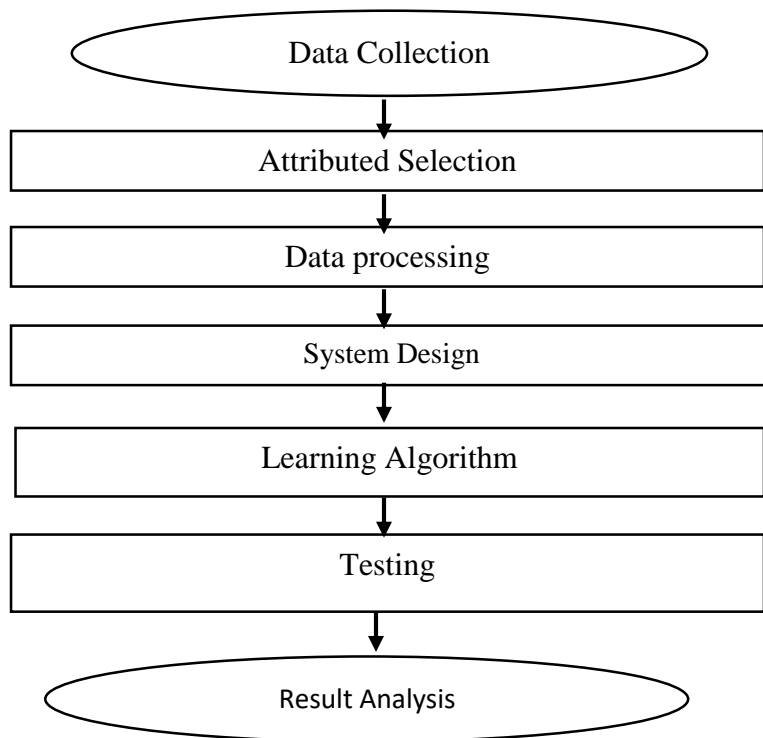


Figure 3.1: Research Design

### **3.3 Data Collection Procedure**

Completing datasets from a variety of sources. Maximum data collection from various sources when combined. This dataset was compiled using several sources from the Liver Patient Dataset. This data collection contains 30691 people's liver information, with the majority being male (72%), female (25%), and others (3 %). I merge the three dataset. I take dataset from Kaggles. This is open source and suitable for any type of research project.

### **3.4 Proposed Methodology**

The liver is in charge of several critical activities in the body, and when it gets unwell or injured, those functions can be lost, causing significant injury to the body. The liver is the biggest organ in the human body, and its prognosis is critical; nonetheless, there has been little study in this area. I am doing research to see if I can come up with the best algorithm to predict liver disease. Some work has already received some interesting research. I used the google colab with python to determine our contribution to this study. In this study, I employ the best method for predicting liver function.

I employ six machine learning prediction algorithms in this research (Logistic Regression, Naive Bayes, SVM, KNN, Decision Tree, and Random Forest). For determining data correctness. In this study, I analyze data accuracy using six machine learning prediction algorithms (Logistic Regression, Naive Bayes, SVM, KNN, Decision Tree, and Random Forest). I use a dataset from Kaggle that is open source and may be used for any type of research project. Performance is compared to previous work to explain the capability of my proposed effort. Let us go over the submission procedure in depth.



### **3.4.1 Data Mining Tool**

The world has grown increasingly digital and technologically dependent. In today's technology age, there are various sorts of data mining tools accessible for data analysis. These materials are all available. When evaluating weather data for forecast, I use kaggle, which is open source and may be used for any type of study. For data analysis, I utilize google colab. I can compute data selection, data preprocessing, preprocessing, visualization, analysis, and classification. Python pandas predicts the likely attribute and possible assumption for projecting liver based on attributes and their connection in the data. Python can accept a variety of data formats as well as numeric and nominal attributes. Python can evaluate data straight from a dataset and respond fast to the results. Using six machine learning prediction algorithms (Naive Bayes, SVM, KNN, Decision Tree, and Random Forest). Google colab in python panda's library allows us to join data in a table and independent applications. Pandas library is an easy-to-use and learn mining program that is regarded as one of the finest in real-time circumstances.

### **3.4.2 Data Processing**

I studied 30691 liver patient data in the research, with 22347 samples being liver patients and 8344 being non-sample patients. The percentage of total liver patients is shown (fig: 3.4 and 3.5). Furthermore, from the liver patient dataset, 21986 male samples and 7806 female samples were taken for analysis (fig: 3.2 and 3.3). This saves time and enables the extraction of information. Deficient, contradictory, and erroneous data are meticulously recorded and deleted throughout preparation. This step yields unmistakably relevant data, which will be used in subsequent processes.

```
df['Gender of the patient'].value_counts()
Male      21986
Female    7803
Name: Gender of the patient, dtype: int64
```

Figure 3.2: Here, Total Number of the male and Female of this data set

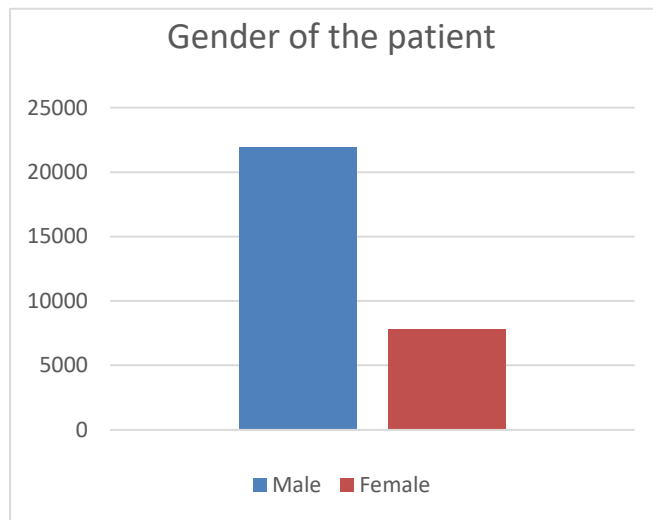


Figure 3.3: Bar chart of Gender of this Dataset

```
df['Result'].value_counts()
1    21917
2     8774
Name: Result, dtype: int64
```

Figure 3.4: Number of the Ratio of patients

Result (It has two category “1” represent for patient already effected by liver disease and ‘2’ represent for patient aren’t effected)

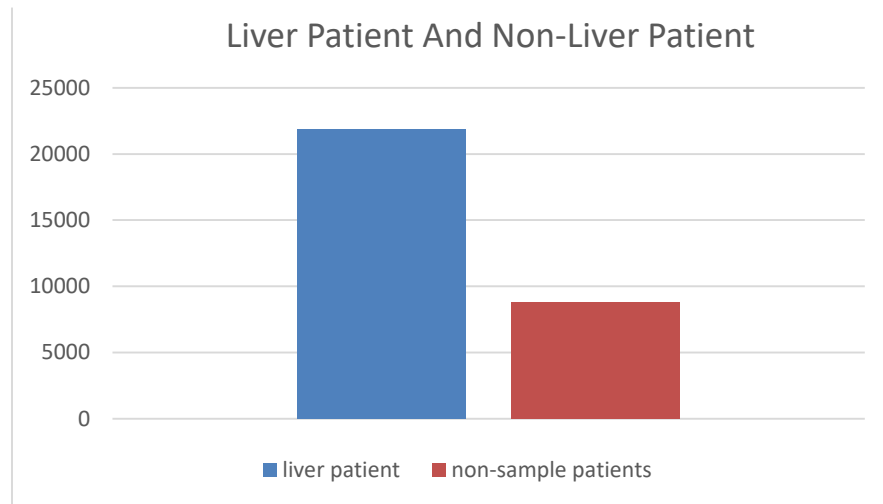


Figure 3.5: Bar chart of liver patient and non-sample patients

### 3.5 System Design

For this analysis, I create a system that will produce the results.

- It has a data batch.
- Prepare dataset for Implementation of classification algorithm
- Feature Selection
- Select attributes (Liver disease dataset)
- Classification algorithm
- Then selected 6 algorithm work at a time
- Then Perform the Performances of Accuracy
- Finally, best technique will find.

System design figure show blow

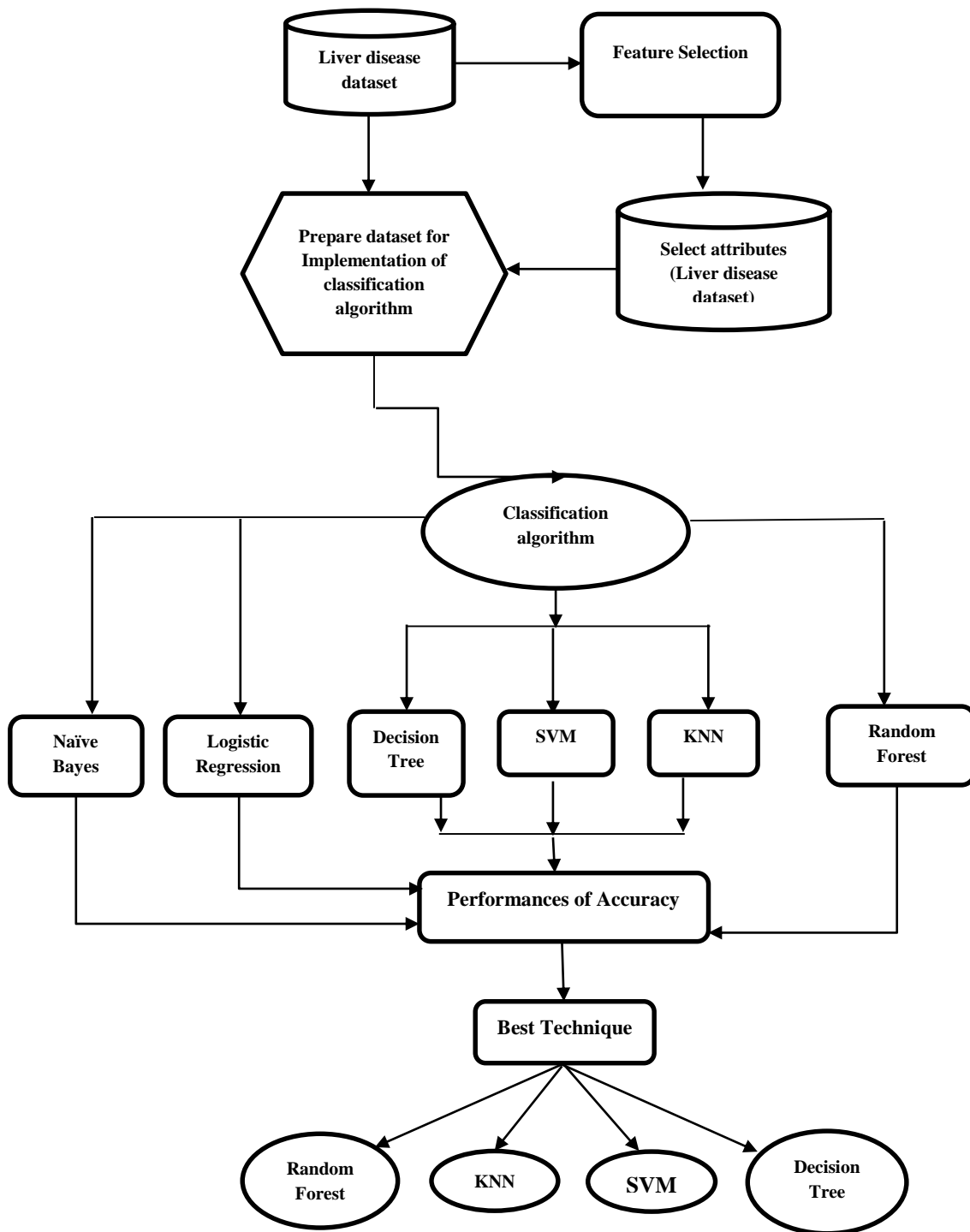


Figure 3.6: System Design

### 3.6 KNN Algorithm

K-Nearest Neighbors Algorithm is the simplest and earliest algorithms that all the available cases stored and classifies the new data measuring similarly. The  $k$  is the number of nearest neighbors considered by the KNN algorithm to take a 'poll' from among these  $K$  Neighbor selection off various values the amount of data points in each category and the outcome of a different classification for the same fundamental item. Displays an example of KNN's work in classifying fresh data. For  $K=4$ , the new data (\*) was classed as 'black.' It was, however, categorized as 'black circle' when  $K=3$ . An image of the K Nearest Neighbor algorithm is simply explained. The value  $K$  is 7 as drawn a circle with sample object here  $K=3$  is as the sample object gets more 'polls' from the black class, it is classified as black circle. Whatever  $k=4$  is the same simple object is classified as black, since the class now get more polls.

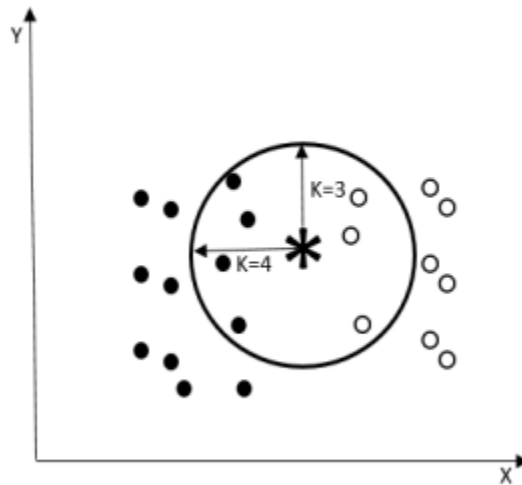


Figure 3.7: KNN Algorithm

### 3.7 Logistic Regression

The algorithm was first proposed by a British Academic Fellow David Cox. Logistic Regression is supervised learning process and classification algorithm of machine learning. Logistic Regression gives dissent value as output and goodness of fit as precision Recall,

accuracy, f1 score, Roc curve, confusion matrix etc. In a word logistic regression developed result in a binary (0/1) format which is used predict the output of categorical.

$$P = \frac{1}{1 + e^{-(b_0 + b_1x_1 + \dots + b_nx_n)}}$$

### 3.8 Decision Tree

Decision Tree measurement is part of the supervised learning algorithm. It is a graphical representation of all possible solutions to a problem that are based on certain criteria and can be easily stated. The benefit of employing a decision tree is that it allows you to create a training model that can be used to forecast the value of the target variable by studying fundamental choice rules gleaned from past data. In Decision Tree, predicting a value label for a commemerate begins at the tree's root. On the basis of equivalence, I compare the root attribute values with the commemerates attribute, the branch is followed in proportion to that value, and I jump to the next node.

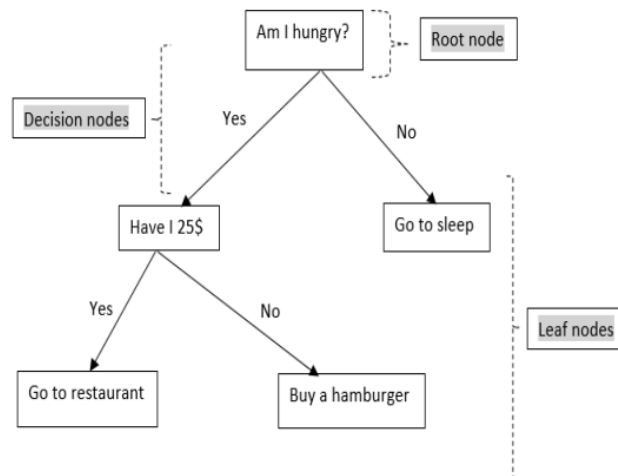


Figure 3.8 Decision Tree

### 3.9 SVM Algorithm

Support vector machine is a type of supervising machine learning that may be used for both classification and regression tasks. Individual observation coordinates are equal to support vectors, and the SVM classification is a frontier that best divides the two groups. Each data item is

represented as a point in n-dimensional space (where n is the number of features), with the value of each feature being the SVM algorithm's value for a certain position.

### **3.10 Naive Bayes**

A Naive Bayes classifier is a simple probabilistic classifier based on the strong independent assumption of the Bayes theorem. A more descriptive definition for model for function would be the underlying probability model's self-determination. In basic words, a Naive Bayes classifier posits that the presence of one class function is independent to the presence of any other feature. Even if the underlying assumption is incorrect, the Naive Bayes classifier performs admirably. The Naive Bayes classifier has the advantage of using only a little quantity of training data to calculate the means and variances of the variables necessary for classification. Because undefined independent variables exist, it is critical to compute only the variances of the variables for each mark rather than the complete covariance matrix. In contrast to the Naive Bayes operator, the Naive Bayes (Kernel) operator may be used to numerical characteristics.

### **3.11 Random Forest Algorithm**

Random forest is a typical classifier that consists of a number of decision trees on different subsets of the provided dataset and takes the average to improve the prediction accuracy of the dataset. The more trees in the forest, the greater the accuracy and the smaller the risk of overfitting. It can also retain accuracy when a significant amount of data is absent. Random can handle both classification and regression task.



### 3.12 Logical Data Model

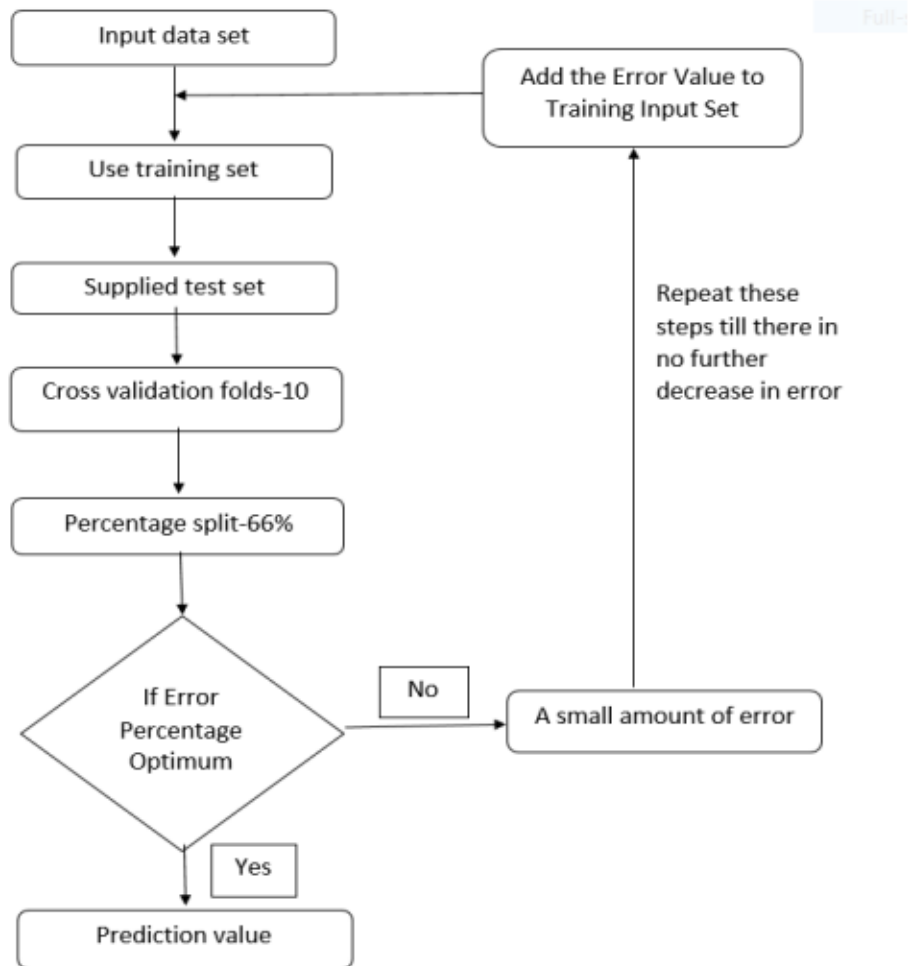


Figure 3.9: Logic Data Model

Firstly, insert data in python panda's library then make data for using training set. After ready training data, ready data for supplied test set. If system percentage error then it replaces it using training set. Then system work again continues.

### 3.13 Algorithm decision Structure

Structure of an Algorithm Part demonstrates how machine learning prediction algorithms function and how the result is obtained. With general, I may begin my work in google colab using a dataset. When and how KNN and Random Forest are used on datasets for liver prediction. KNN and Random Forest analyze data based on the data structure and information provided in the dataset. Google colab has various rules and relationships, as well as a tree structure. There are six types of predictions based on data, including successfully categorized instances, erroneously classified instances, recall, accuracy, specificity, sensitivity, and F1 measure. The algorithm demonstrates a relationship between correctly classified occurrences, incorrectly classed instances, recall, accuracy, specificity, sensitivity, and the F1 measure. The first row of the graph in figure [8] depicts the relationship between correctly classified occurrences, incorrectly classified instances, recall, accuracy, specificity, sensitivity, and F1 measure. Random forest is a typical classifier that consists of a number of decision trees on different subsets of the provided dataset and takes the average to improve the prediction accuracy of the dataset. The more trees in the forest, the greater the accuracy and the smaller the risk of overfitting. It can also maintain accuracy when a large amount of data is missing.

### 3.14 Pair plot graph between each Entities

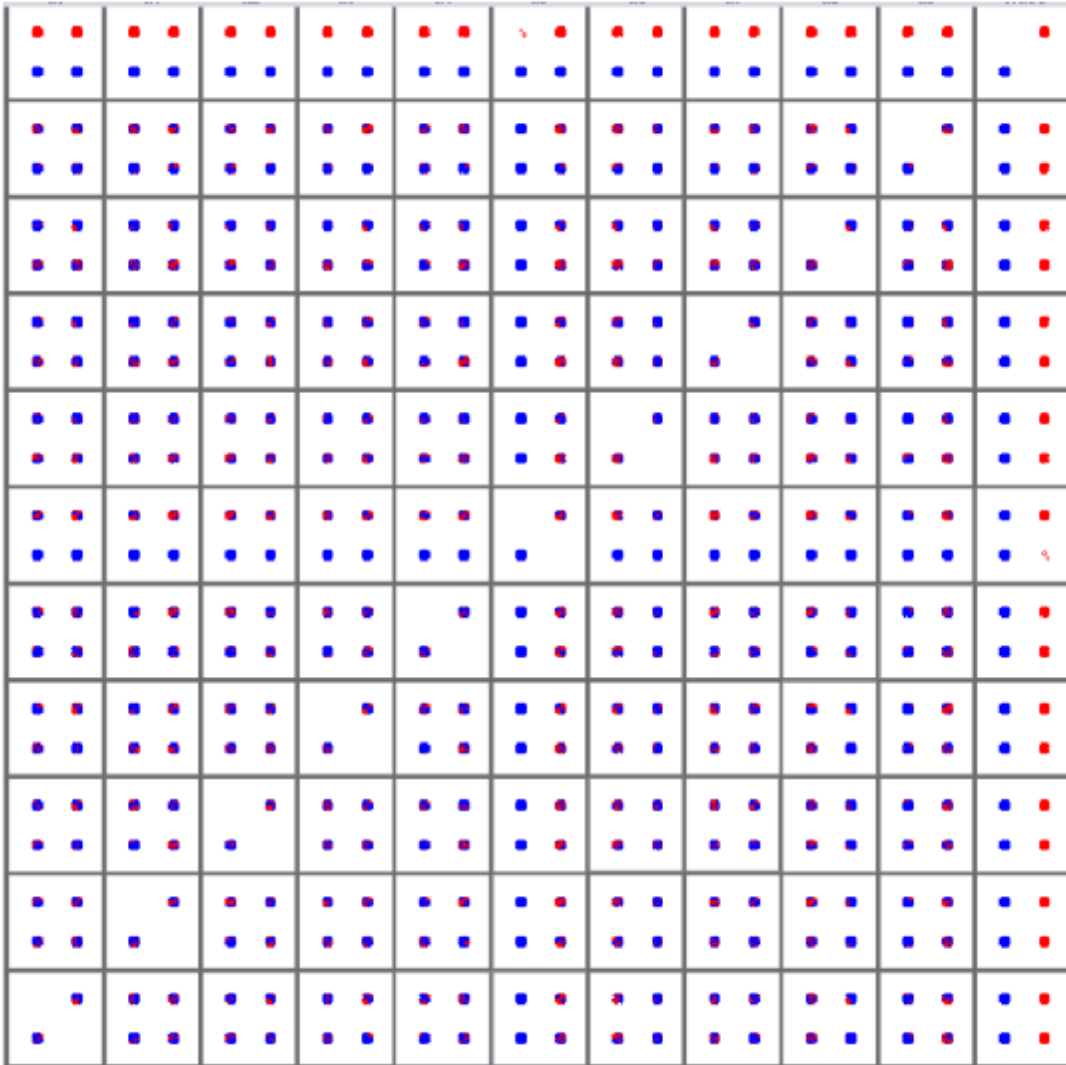


Figure 3.10: Pair plot graph between each Entities

To begin, I need to know what data I have. I can enter the socioeconomic data into Pandas as a data frame and examine the columns:

	Age of the patient	Gender of the patient	Total Bilirubin	Direct Bilirubin	Alkphos Alkaline Phosphotase	Sgpt Alamine Aminotransferase	Sgot Aspartate Aminotransferase	Total Protiens	ALB Albumin	A/G Ratio Albumin and Globulin Ratio	Result
0	65.0	Female	0.7	0.1	187.0	16.0	18.0	6.8	3.3	0.90	1
1	62.0	Male	10.9	5.5	699.0	64.0	100.0	7.5	3.2	0.74	1
2	62.0	Male	7.3	4.1	490.0	60.0	68.0	7.0	3.3	0.89	1
3	58.0	Male	1.0	0.4	182.0	14.0	20.0	6.8	3.4	1.00	1
4	72.0	Male	3.9	2.0	195.0	27.0	59.0	7.3	2.4	0.40	1
5	46.0	Male	1.8	0.7	208.0	19.0	14.0	7.6	4.4	1.30	1
6	26.0	Female	0.9	0.2	154.0	NaN	12.0	7.0	3.5	1.00	1
7	29.0	Female	0.9	0.3	202.0	14.0	11.0	6.7	3.6	1.10	1
8	17.0	Male	0.9	0.3	202.0	22.0	19.0	7.4	4.1	1.20	2
9	55.0	Male	0.7	0.2	290.0	53.0	58.0	6.8	3.4	1.00	1

Figure 3.11: Dataset index

Each row of information addresses one perception for one segment, which brings about one section addressing the factors (information in this organization is known as clean information). There are two classes of sections: one for crude information perceptions and one more for perception results, which show which patients have liver infections and which don't. In the outcome segment, '2' signifies patients who are not impacted by liver sickness and '1' means patients who are impacted by liver illness.

## CHAPTER 4

### Implementation and Result Analysis

#### 4.1 Introduction

In this study, I chose machine learning-based algorithm to determine the data set's prediction accuracy. In the work, I used some factual estimations to determine how well different classification algorithms worked in tests. I take our data from the Liver Patient Dataset and concentrate on the dataset's 11 attributes to assess the accuracy of our predictions. Here I worked with 7 machine learning algorithms, in which I find out the best accuracy from KNN (K nearest Neighbor), and random forest algorithm.

#### 4.2 Six Machine Learning Prediction Algorithms Applying Procedure

First, I need to select the desired data set to get the best accuracy. Then I have to fill all missing data using mean technique in python panda's library. Then I have to generate this file on google colab. Our data set have 30691 Patients data. In this dataset male 72 percent, female 25 percent and others 3 percent. This study made use of classification algorithms, such as Naive Bayes, LR (Logistic Regression), SVM (Support Vector Machine), KNN (K nearest Neighbor), Decision Tree and Random Forest for liver disease prediction. In this case, I trained 80% of the data and tested 20% of it before applying six algorithms one by one. Determine the confusion matrix for each algorithm like TP (True Positive), FP (False Positive), FN (False Negative), and TN (True Negative) then determine accuracy ( $\frac{TP+TN}{TP+FP+TN+FN}$ ), precision( $\frac{TP}{TP+FP}$ ), recall( $\frac{TP}{TP+FN}$ ), f1-score( $\frac{2*Precision*Recall}{Precision+Recall}$ ). Represent visualization between actual value and predict value. I obtain the greatest accuracy in KNN (K nearest Neighbor), Random Forest and Decision Tree after using this strategy.

#### 4.3 Experimental Result and Analysis

I utilized numerous experiments in this experiment to evaluate the six-machine learning classifier on the liver disease dataset. To classify the various results, the following steps were utilized.

### 4.3.1 Naïve Bayes

Here, Test data was 20% and train data was 80% among 30691 data. Then determined confusion matrix.

TABLE 4.1: CONFUSION MATRIX FOR NAIVE BAYES

	True	False
Positive	1806	2635
Negative	56	1633

In this case, TP = 1806, FP = 2635, TN = 56 and FN = 1633

Now, the classification report

TABLE 4.2: CLASSIFICATION REPORT FOR NAIVE BAYES

	Precision	Recall	F1-score	Support
1	0.97	0.41	0.57	4441
2	0.38	0.96	0.55	1698
Accuracy			0.56	6139
Macro Avg.	0.67	0.68	0.56	6139
Weighted Avg.	0.80	0.56	0.57	6139

Now, Visualization between actual value and predicted value.

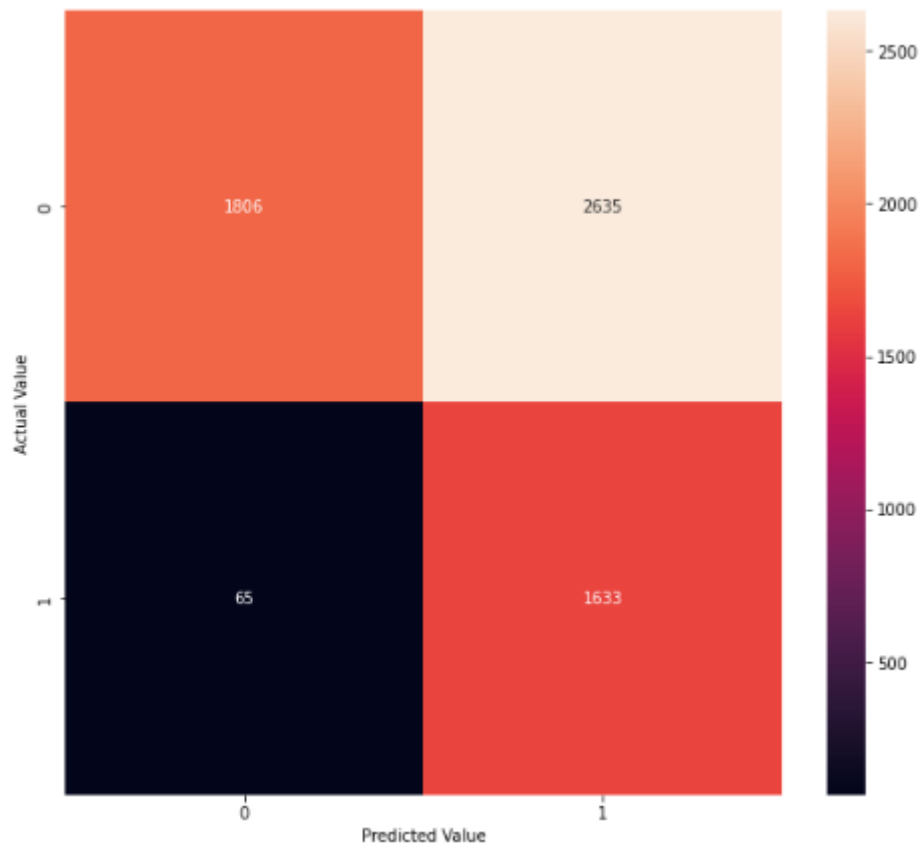


Figure 4.1: Visualization between actual value and predicted value for Naïve Bayes

Then, prediction accuracy is 56% for Naïve Bayes.

### 4.3.2 Logistic Regression

Here, Test data was 20% and train data was 80% among 30691 data. Then determined confusion matrix.

TABLE 4.3: CONFUSION MATRIX FOR LOGISTIC REGRESSION

	True	False
Positive	4152	289
Negative	1447	251

In this case, TP = 4152, FP = 289, TN = 1447 and FN = 251

Now, the classification report

TABLE 4.4: CLASSIFICATION REPORT FOR LOGISTIC REGRESSION

	Precision	Recall	F1-score	Support
1	0.74	0.93	0.83	4441
2	0.46	0.15	0.22	1698
Accuracy			0.72	6139
Macro Avg.	0.60	0.54	0.53	6139
Weighted Avg.	0.67	0.72	0.66	6139



Now, Visualization between actual value and predicted value.

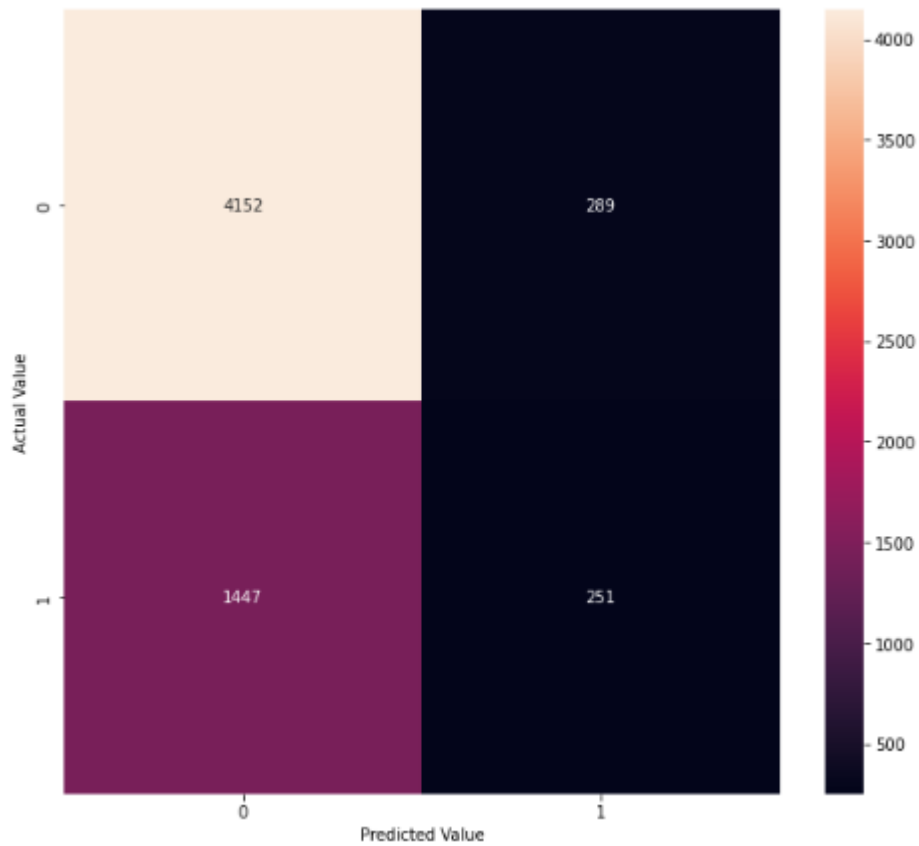


Figure 4.2: Visualization between actual value and predicted value Logistic Regression

Then, prediction accuracy is 72% for Logistic Regression.

### 4.3.3 Decision Tree

Here, Test data was 20% and train data was 80% among 30691 data. Then determined confusion matrix.

TABLE 4.5: CONFUSION MATRIX FOR DECISION TREE

	True	False
Positive	4415	26
Negative	34	1664

In this case, TP = 4412, FP = 29, TN = 35 and FN = 1663

Now, the classification report

TABLE 4.6: CLASSIFICATION REPORT FOR DECISION TREE

	Precision	Recall	F1-score	Support
1	0.99	0.99	0.99	4441
2	0.46	0.98	0.98	1698
Accuracy			0.99	6139
Macro Avg.	0.99	0.99	0.99	6139
Weighted Avg.	0.99	0.99	0.99	6139

Now, Visualization between actual value and predicted value.

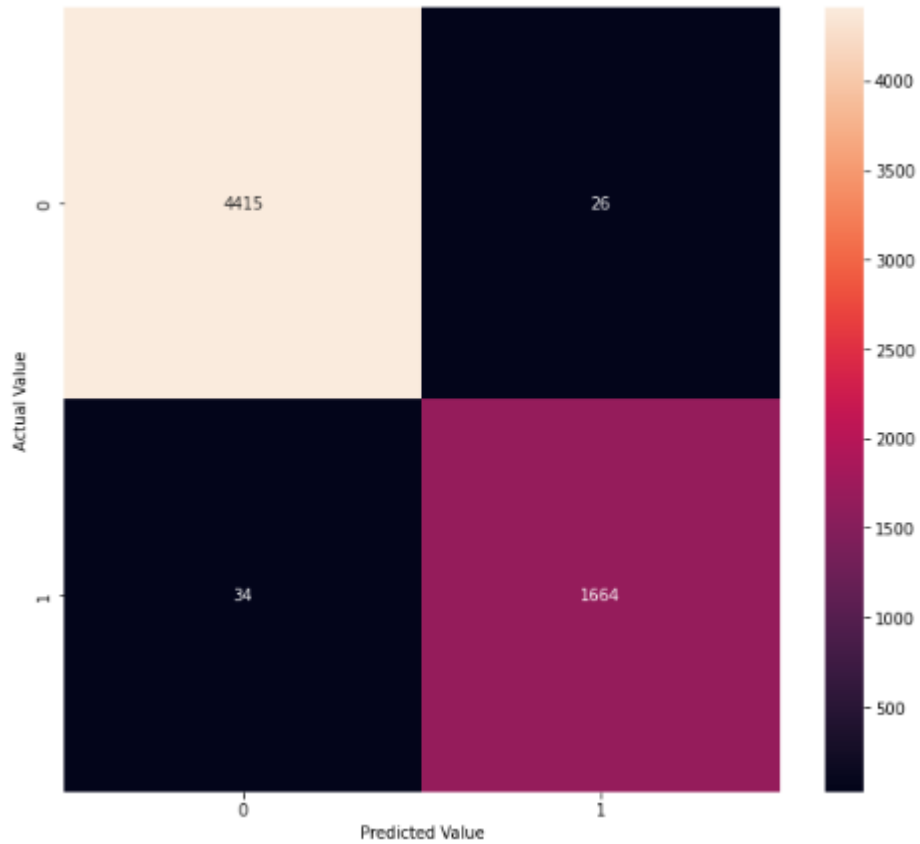


Figure 4.3: Visualization between actual value and predicted value Decision Tree

Then, prediction accuracy is 99% for Decision Tree.

#### 4.3.4 SVM

Here, Test data was 20% and train data was 80% among 30691 data. Then determined confusion matrix.

TABLE 4.7: CONFUSION MATRIX FOR SVM

	True	False
Positive	4429	12
Negative	134	1564

In this case, TP = 4412, FP = 29, TN = 35 and FN = 1663

Now, the classification report

TABLE 4.8: CLASSIFICATION REPORT FOR SVM

	Precision	Recall	F1-score	Support
1	0.97	1.00	0.98	4441
2	0.99	0.92	0.96	1698
Accuracy			0.97	6139
Macro Avg.	0.98	0.96	0.97	6139
Weighted Avg.	0.98	0.98	0.98	6139

Now, Visualization between actual value and predicted value.

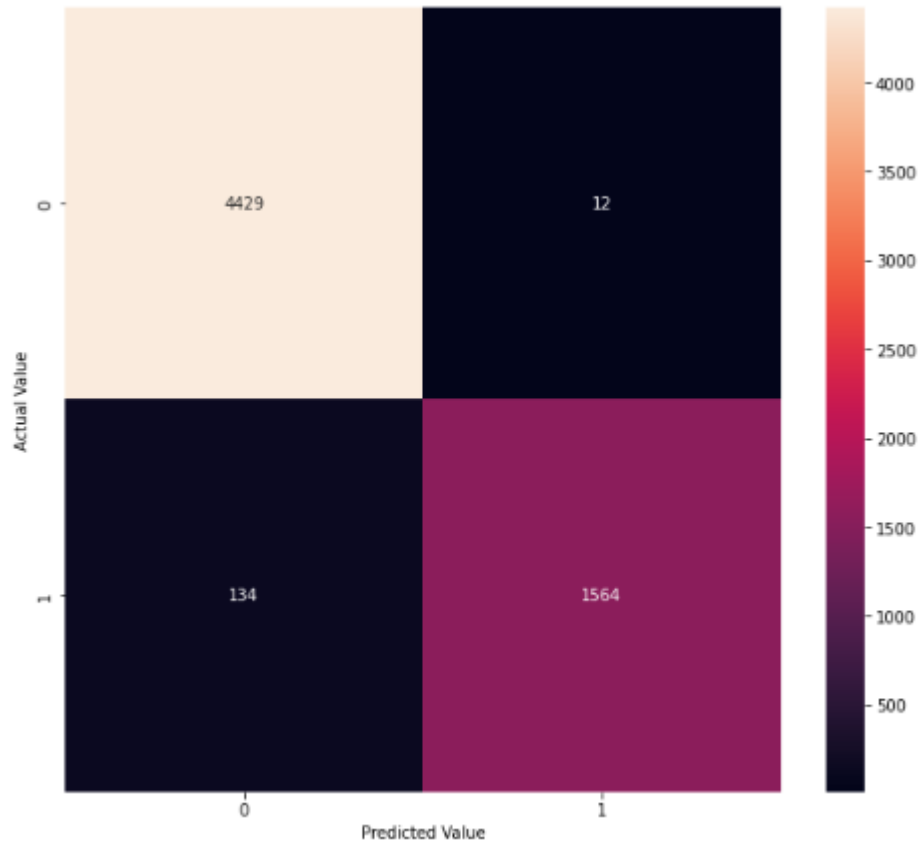


Figure 4.4: Visualization between actual value and predicted value for SVM

Then, prediction accuracy is 97% for SVM.

### 4.3.5 KNN

Here, Test data was 20% and train data was 80% among 30691 data. Then determined confusion matrix.

TABLE 4.9: CONFUSION MATRIX FOR KNN

	True	False
Positive	4335	106
Negative	131	1567

In this case, TP = 4335, FP = 106, TN = 131 and FN = 1567

Now, the classification report

TABLE 4.10: CLASSIFICATION REPORT FOR KNN

	Precision	Recall	F1-score	Support
1	0.97	0.98	0.97	4441
2	0.94	0.92	0.93	1698
Accuracy			0.96	6139
Macro Avg.	0.95	0.95	0.95	6139
Weighted Avg.	0.96	0.96	0.96	6139

Now, Visualization between actual value and predicted value.

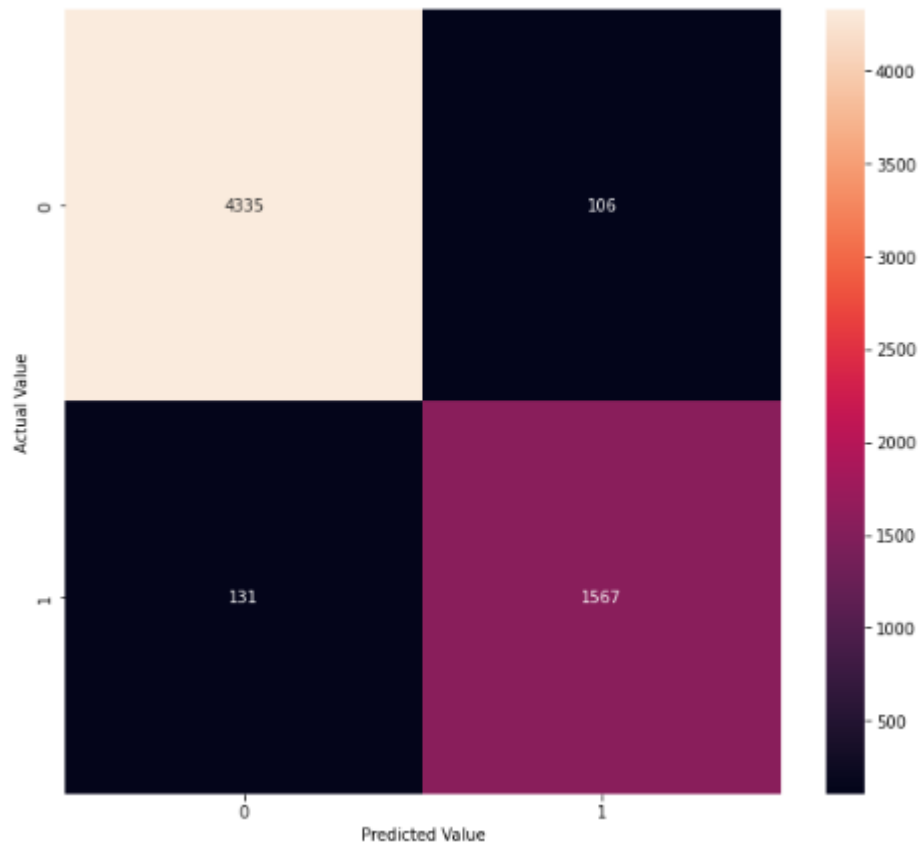


Figure 4.5: Visualization between actual value and predicted value for KNN

Then, prediction accuracy is 96% for KNN.

### 4.3.6 Random Forest

Here, Test data was 20% and train data was 80% among 30691 data. Then determined confusion matrix.

TABLE 4.11: CONFUSION MATRIX FOR RANDOM FOREST

	True	False
Positive	4432	9
Negative	22	1676

In this case, TP = 4432, FP = 9, TN = 22 and FN = 1676

Now, the classification report

TABLE 4.12: CLASSIFICATION REPORT FOR RANDOM FORESRT

	Precision	Recall	F1-score	Support
1	1.00	1.00	1.00	4441
2	0.99	0.99	0.99	1698
Accuracy			0.99	6139
Macro Avg.	0.99	0.99	0.99	6139
Weighted Avg.	0.99	0.99	0.99	6139



Now, Visualization between actual value and predicted value.

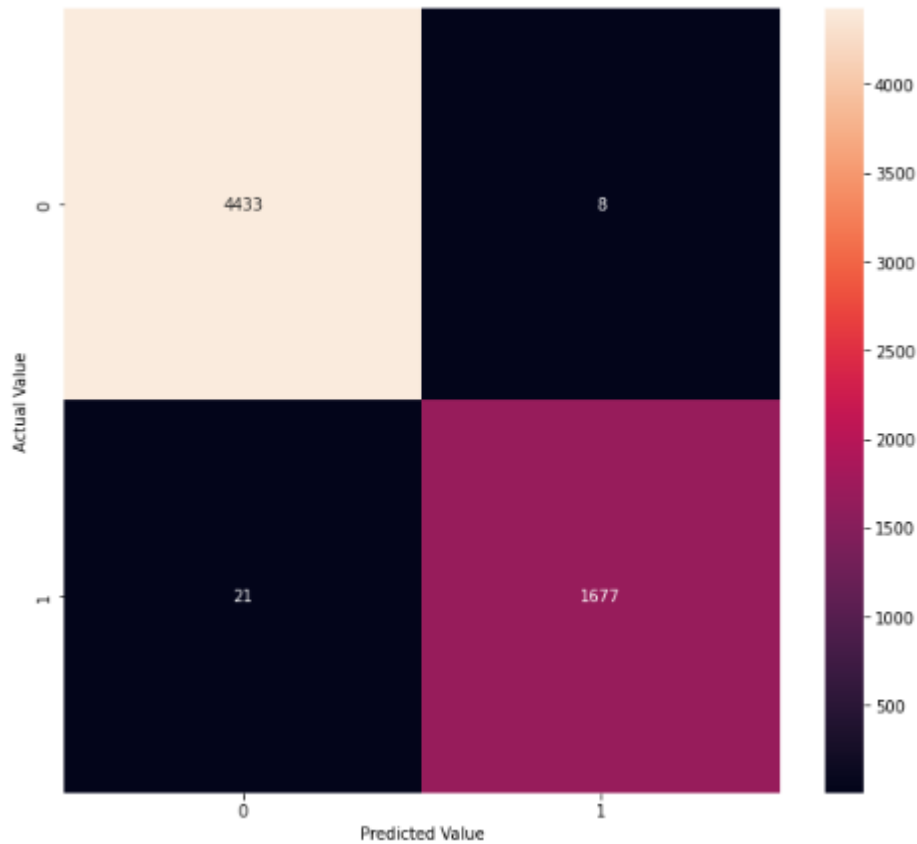


Figure 4.6: Visualization between actual value and predicted value for Random Forest

Then, prediction accuracy is 96% for Random Forest.

#### 4.4 Result Analysis

To classify the various results, the following definitions were utilized. First, the outcomes of different classifiers are compared in terms of successfully classified instances (fig. 1) using the training set, test set, and accuracy percentage indicated in the table.

TABLE 4.13: ACCURACY MEASURE

Algorithms	Precision	Recall	F1-score	Accuracy
Naïve Bayes	0.97	0.41	0.57	56%
Logistic Regression	0.74	0.93	0.83	72%
Decision Tree	0.99	0.99	0.99	99%

SVM	0.97	1.00	0.95	98%
KNN	0.97	0.98	0.97	96%
Random Forest	1.00	1.00	1.00	99%

The highest accuracy is achieved by four algorithms: Decision Tree (99%), SVM (98%) KNN (96%), and Random Forest (99%).

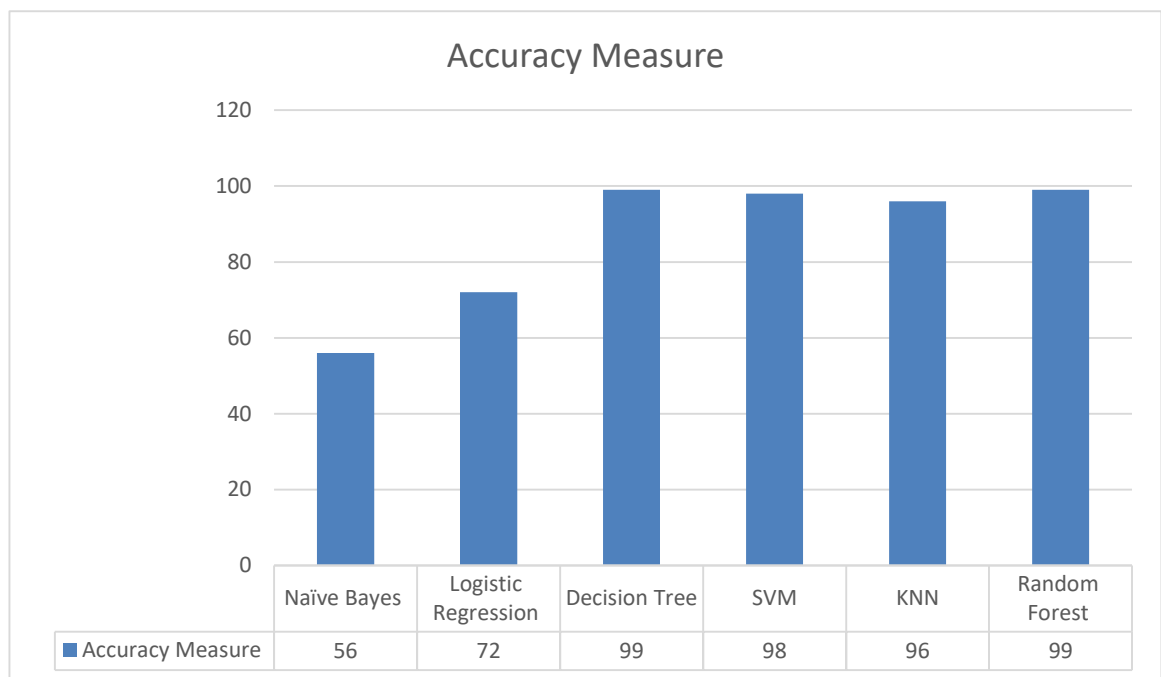


Figure 4.7: Accuracy Measure Visualization

#### 4.5 Description of My Work

In this method, I employ the updated version of Direct Relapse to calculate the liver prediction estimation. The following advancements exemplify the technique's method:

- I collect the knowledge from the beginning and then review it. I am making a six month-long deal about how to do our job.
- The data instructive lists detail the readiness and test data.

- The dataset would then be summed up.
- Creating a mental image of the data to demonstrate its present state.
- Here I have used 6 machine learning based algorithms.
- At the moment I am going to apply that algorithm to google colab to get our desired accuracy.
- Here I apply use precision, recall, f1-score, support, and Accuracy options
- Authorizing this dataset, collection into two parts preparation and testing.

In google colab the highest accuracy is achieved by four algorithms: Decision Tree (99%), SVM (98%) KNN (96%), and Random Forest (99%). And other algorithm shows Naïve Bays 56 percent and Logistic Regression 72 percent accuracy.

The machine gaining knowledge of techniques said in this research can help medical practitioners but aren't an alternative for making judgments based totally on ML classifiers for diagnostic paths. Many of the restrictions that stand up in healthcare are due to inaccuracy in analysis, missing records, price, and time. By enhancing detection of danger elements and diagnostic variables, ML strategies can help minimize the worldwide burden of liver disorder on public fitness. More appreciably, for continual liver contamination, the use of ML to detect liver ailment at an earlier level or in hid times may lessen liver-related mortality, transplants, and/or hospitalizations. Early identity improves analysis since therapy can start earlier than the disease progresses to later stages. Invasive diagnostics, along with biopsy, might also be less commonplace in this circumstance. Although this work focused on hepatitis and chronic liver disease traits for ML training, its miles feasible that the methods can be utilized to distinguish between different types of liver disorder and healthful people. Applying all the techniques mentioned above to different fields of medicine might pave the way for AI or ML-assisted diagnostics. This research's relaxation is organized as follows. The second section analyzes all studies on liver infection. The strategies used to fulfill the data-mining challenge are described in Section 3. Section 4

addresses the findings in relation to the dataset and the assessment criteria, and Section 5 concludes with a few ideas for similar investigations.

## CHAPTER 5

### Conclusion and Future Work

#### 5.1 Conclusion

In the proposed review, various classifiers had been applied to a dataset of liver patient problems to foresee liver diseases through the utilization of made programming program. The dataset changed into handled and did the use of the google colab gadget. Procedures for determination the utilization of a 10-overlay move approval evaluating elective as far as execution time and precision, the proposed works of art's results have been in examination the utilization of and without trademark choice methods.

Any abnormality in liver function that produces sickness is referred to as liver disease. The liver is in charge of various critical actions in the body, and if it becomes unwell or damaged, these functions can be jeopardized, resulting in serious bodily harm. Hepatic disease is another word for liver disease. The phrase "liver disease" refers to any condition that causes the liver to fail to perform its intended function. Before there is a decline in function, more than 75%, or three quarters, of the liver tissue must be damaged. Excessive acetaminophen and acetaminophen combination drugs like Vicodin and Norco, as well as statins, cirrhosis, alcoholism, hepatitis A, B, C, D, and E, infectious mononucleosis (Epstein Barr virus), non-alcoholic fatty liver disease (NASH), and iron overload can all affect the liver (hemochromatosis). Common signs of liver illness include nausea, vomiting, right upper quadrant stomach pain, and jaundice (a yellow discoloration of the skin due to elevated bilirubin concentrations in the bloodstream). Tiredness, lethargy, and weight loss are all possible side effects. However, because there are so many different types of liver diseases, most of the symptoms are unique to one disease until it progresses to late-stage liver disease and liver failure. In Bangladesh, one out of every three people has liver disease. More than 40 million Bangladeshis suffer from fatty liver disease, and 10 million have untreated hepatitis B and C. In all, 59,227 patients were included in the report (age ranged 15-95 years). Men make up the vast majority of patients (67.9 percent). While all patients who visited the hepatology clinic had liver problems, the vast majority (77.35

percent) had nonnuclear dyspepsia or irritable bowel syndrome. Chronic liver diseases (CLDs) are the most frequent kind of liver disease in people (37 -69 percent). In countries with a more strong healthcare system, hepatic encephalopathy, a complication of CLD, was less prevalent. Other infections, such as liver abscess and biliary ascariasis, were widespread. Another prevalent ailment was acute hepatitis, which accounted for approximately 20% of all cases. The study included 30691 individuals who presented to hepatology departments at various medical institutions. Patients make up 72 percent of the group, 25 percent are female, and the remaining 3 percent are male. They ranged in age from 15 to 95 years old. Individuals under the age of 15 were not eligible to participate. Males make up the bulk of the patients. Our study's male predominance is unsurprising. Males are more likely to develop HBV-related CLD, such as liver cirrhosis. This is especially true for males over the age of 40, and our patients' average age was 51.95. Similarly, men are more likely to develop hepatocellular carcinoma (HCC), with HBV accounting for 61.5 percent of HCC cases in Bangladesh. In the Dhaka division, CLD was utilized by 37% of patients. The percentages are similar in the divisions of Barisal (38%) and Khulna (39%), both of which are located in southern Bangladesh. However, only 22.8 percent of patients in the Sylhet division had CLD, while the numbers in Chittagong and Rajshahi divisions were exceedingly high, at 50 and 69 percent, respectively. Many people go to their liver treatment late without realizing it, which causes a lot of financial and physical damage to them. For this reason they can use it to prevent this problem. They find out the best accuracy of the liver so they can look for medical help before they have more liver problems. To boost and discover liver prognosis Accuracy is required for the training accreditation training phase. A decision-making mechanism with several standards that will limit the number of accreditation changes. The procedure analyzes internal and external parameters and prioritizes the coefficients used to make an educated judgment using the instrument. It would also aid the authority in the creation of creative educational programs aimed at identifying the significance of different decision elements. This approach employs a normalized protocol to evaluate individual measures before combining them to produce the final forecast. For the same data, it generates different algorithm

results. As a consequence, the difficulties of using an additive approach to estimate liver function are resolved by using a multiplicative approach with a classification function. The implications of big data research on liver disease are explored in this article. I applied our proposed 7 machine learning based algorithm to the liver patient's dataset, which was taken from an open source Kaggle liver patient's dataset from the authority Kaggle database, for perfect liver disease determination. The decision to build and expand a tree model based on similar data is in accordance with tree standards. To improve the discovery perspective, Kuggle, a high-level data mining device that facilitates direct access from data sets, is employed. This method is ideal for evaluating reliable data. The obtained results outperform the current approach, showing the accuracy of the expectation via the straight relapse estimation. Chronicled data were used primarily for the analysis of liver disease details in this work [12] the impact of liver illness is great in Bangladesh, but it exists elsewhere. With the purpose of predicting liver illness later on, understanding the causes of liver disease is highly useful for those individuals who go to their liver treatment late without recognizing it, causing them a lot of financial and bodily harm [13]. The hepatic disorder is any other call for liver ailment. Symptoms of hepatic disorders include nausea, vomiting, exhaustion, belly ache, swelling, again ache, lethargy, and weight reduction. Certain individuals had been found to have jaundice (yellowing of the skin and eyes), fluid in an odd hollow space, light feces, and, specifically, an enlarged spleen and gallbladder. Imaging studies and liver function tests can locate liver damage and useful resource in the prognosis of acute and continual liver issues. The acute liver disorder is described as having a period of much less than six months with a history of the illness.

It delivers numerous calculation results for similar information. Subsequently, the issues related to utilizing an added substance method to gauge liver capability are overwhelmed by utilizing a multiplicative methodology with a grouping capability. This article explores the ramifications of large information research on the liver ailment. For perfect liver sickness evaluation, I executed our recommended 7 machine learning-based techniques to the liver patient's dataset, which was gathered from an open-source Kaggle liver patient's

dataset from the legitimate Kaggle data set. The decision to create and grow a tree model based on comparable information is predictable with tree norms.

## **5.2 Future Work**

I utilize liver patient data to forecast with the highest accuracy possible. I utilize the patient's age, gender, Total Bilirubin, Direct Bilirubin, Alkphos Alkaline Phosphotase, Sgpt Alamine Aminotransferase, Sgot Aspartate Aminotransferase, Total Proteins, ALB Albumin, A/G Ratio Albumin, Globulin Ratio, and kaggle results. I use logistic regression, naive bays, SMV, KNN, decision tree and random forest etc ; From which I get the best accuracy to random forest and KNN algorithm. After that, I may compare which strategy is best for our data. In the future, I can improve our system by include many more factors. (Hepatitis B (HBV) and C (HCV), non-alcoholic steatohepatitis (NASH), malnutrition, toxins, and several tropical illnesses) to improve the accuracy of our liver disease forecasting software and initiatives [14]. Artificial intelligence may have a future in liver prediction. In recent years, liver prediction has advanced tremendously. AI can help liver doctors enhance their capacity to forecast the weather in the future. In the future, the local interpretable model-agnostic explanation (LIME) technique will be used to understand the model's interpretability. Instead of binary classification, multinomial classification can be used to distinguish between different types of liver disease. The performance of each model can thus be compared. The ML algorithms discussed here can help health sectors obtain better diagnosis by finding groups or levels within medical data to help healthcare practitioners. Furthermore, ML approaches are data-driven, and they employ diagnostic information directly from patients' medical testing. As a result, it is a more dependable process. The ML methods used in this article can save time, money, and potentially lives by improving disease diagnosis. At the point when liver hindrance continues to a high-level certificate, liquid collects inside the legs (edema) and the stomach (ascites). Ascites can reason bacterial peritonitis, a without a doubt destructive disease. A singular will wound or drain essentially on the off chance that the liver postponements or stops the creation of the proteins expected for blood coagulation.



In future, I work with online software batch Liver Disease prediction including NLP and AI with image processing also observe the patient behavioral fact.

## REFERENCES

- [1] J. Jacob, J. C. Mathew, J. Mathew, and E. Issac, "Diagnosis of liver disease using machine learning techniques," *Int Res J Eng Technol*, vol. 5, no. 04, 2018.
- [2] E. A. El-Shafeiy, A. I. El-Desouky, and S. M. Elghamrawy, "Prediction of liver diseases based on machine learning technique for big data," in *International conference on advanced machine learning technologies and applications*. Springer, 2018, pp. 362–374.
- [3] S. Vijayarani and S. Dhayanand, "Liver disease prediction using svm and naïve bayes algorithms," *International Journal of Science, Engineering and Technology Research (IJSETR)*, vol. 4, no. 4, pp. 816–820, 2015.
- [4] A. S. Rahman, F. J. M. Shamrat, Z. Tasnim, J. Roy, and S. A. Hossain, "A comparative study on liver disease prediction using supervised machine learning algorithms," *International Journal of Scientific & Technology Research*, vol. 8, no. 11, pp. 419–422, 2019.
- [5] S. R. Ghosh and S. Waheed, "Analysis of classification algorithms for liver disease diagnosis," *Journal of Science Technology and Environment Informatics*, vol. 5, no. 1, pp. 360–370, 2017.
- [6] Y. Kumar and G. Sahoo, "Prediction of different types of liver diseases using rule based classification model," *Technology and Health Care*, vol. 21, no. 5, pp. 417–432, 2013.
- [7] E. M. Hashem and M. S. Mabrouk, "A study of support vector machine algorithm for liver disease diagnosis," *American Journal of Intelligent Systems*, vol. 4, no. 1, pp. 9–14, 2014.
- [8] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC medical informatics and decision making*, vol. 19, no. 1, pp. 1–16, 2019.

- [9] J. Singh, S. Bagga, and R. Kaur, "Software-based prediction of liver disease with feature selection and classification techniques," *Procedia Computer Science*, vol. 167, pp. 1970–1980, 2020.
- [10] S. Naggie, S. Lusk, J. W. Thompson, M. Mock, C. Moylan, J. E. Lucas, L. Dubois, L. St John-Williams, M. A. Moseley, and K. Patel, "Metabolomic signature as a predictor of liver disease events in patients with hiv/hcv coinfection," *The Journal of Infectious Diseases*, vol. 222, no. 12, pp. 2012–2020, 2020.
- [11] V. Vats, L. Zhang, S. Chatterjee, S. Ahmed, E. Enziama, and K. Tepe, "A comparative analysis of unsupervised machine techniques for liver disease prediction," in *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE, 2018, pp. 486–489.
- [12] M. B. Priya, P. L. Juliet, and P. Tamilselvi, "Performance analysis of liver disease prediction using machine learning algorithms," *International Research Journal of Engineering and Technology (IRJET)*, vol. 5, no. 1, pp. 206–211, 2018.
- [13] C.-C. Wu, W.-C. Yeh, W.-D. Hsu, M. M. Islam, P. A. A. Nguyen, T. N. Poly, Y.-C. Wang, H.-C. Yang, and Y.-C. J. Li, "Prediction of fatty liver disease using machine learning algorithms," *Computer methods and programs in biomedicine*, vol. 170, pp. 23–29, 2019.
- [14] M. S. Borah, B. P. Bhuyan, M. S. Pathak, and P. Bhattacharya, "Machine learning in predicting hemoglobin variants," *Int J Mach Learn Comput*, vol. 8, no. 2, pp. 140–3, 2018.
- [15] T. Hydes, W. Gilmore, N. Sheron, and I. Gilmore, "Treating alcoholrelated liver disease from a public health perspective," *Journal of hepatology*, vol. 70, no. 2, pp. 223–236, 2019.
- [16] Nahar N, Ara F. "Liver disease prediction by using different decision tree techniques", *International Journal of Data Mining & Knowledge Management Process*, vol 8, no.2, pp. 1-9, 2018.
- [17] Rajeswari P, Reena GS. "Analysis of liver disorder using data mining algorithm", *Global journal of computer science and technology*, 2010.

- [18] Durai V, Ramesh S, Kalthireddy D. “Liver disease prediction using machine learning”, *Int. J. Adv. Res. Ideas Innov. Technol*, vol 5, no.2, pp. 1584-8, 2019.
- [19] Yao Z, Li J, Guan Z, Ye Y, Chen Y. “Liver disease screening based on densely connected deep neural networks. Neural Networks.”, *International Journal of Data Mining & Knowledge Management Process*, vol 8, no.2, pp. 123-299, 2020.
- [20] Wang NC, Zhang P, Tapper EB, Saini S, Wang SC, Su GL. “Automated measurements of muscle mass using deep learning can predict clinical outcomes in patients with liver disease”, *The American journal of gastroenterology*, vol 115, no.2, pp. 1210, 2020.

Ts/A/riqes

## Arafatur Rahman Soikat(221-25-111)\_M.Sc

### ORIGINALITY REPORT

19%

SIMILARITY INDEX

12%

INTERNET SOURCES

13%

PUBLICATIONS

11%

STUDENT PAPERS

### PRIMARY SOURCES

- 1** Fahad Mostafa, Easin Hasan, Morgan Williamson, Hafiz Khan. "Statistical Machine Learning Approaches to Liver Disease Prediction", Livers, 2021 **1%**

Publication
- 2** Salimur Rahman, Mohammad Faroque Ahmed, Mohammad Jamshed Alam, Mohammad Izazul Hoque, Chitta Ranjan Debnath. "Distribution of Liver Disease in Bangladesh: A Cross-country Study", Euroasian Journal of Hepato-Gastroenterology, 2014 **1%**

Publication
- 3** Sai Rohith Tanuku, Addike Ajay Kumar, Sai Roop Somaraju, Rushitaa Dattuluri, Madana Vamshi Krishna Reddy, Sambhav Jain. "Liver Disease Prediction Using Ensemble Technique", 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS), 2022 **1%**

Publication