Final Year Project Report

**Full Unit: Final Report**



Project Title: liver disease prediction system using machine learining

Student’s first and last name

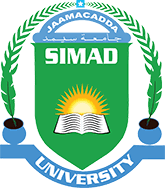
Abdirhamn ismail elmi

Alisaid adan Hassan

A report submitted in part fulfilment of the degree of

**BIT28 Information Technology**

**Supervisor:** Abdikarim Abdi Abulle



Department of Information Technology

Faculty of computing, Simad University

Augusto ,2025

# DECLARATION.A

“I declare that the following is my own work and does not contain any unacknowledged work from any other sources. This project was undertaken to fulfill the requirements of the bachelor’s degree program in Computer Science/Information Technology/Graphics, and Multimedia at Simad University”.

Name: Alisaid adan Hassan

Student Signature : ……………………………

Name: Abdirhamn ismail elmi

Student Signature : ……………………………

Date : ……………………….……

DECLARATION. B

Eng. Abdikarim Abdi Abulle

ACKNOWLEDGEMENTS

**Student 1:** Alisaid adan Hassan

I am very thankful my supervisor **Eng. Abdikarim Abdi Abulle** for his support and guidance. He was not only my supervisor but also gave me valuable ideas that improved this project. His advice and encouragement helped me complete this research. also thank the **faculty of Computing at SIMAD University** for their technical support and for allowing me to use the laboratory facilities. I am grateful to my teammates, especially **Abdirahman ismail elmi,** for their teamwork and dedication, which made this journey more meaningful. I am also deeply thankful to my family and friends for their love, support, and encouragement. They gave me the strength and motivation to finish my work. Finally, I appreciate everyone who helped me in any way, including my classmates and study participants. Your support made this research possible, and I truly appreciate it.

**Student 2:** Abdirhamn ismail elmi

Also, I am very grateful to **Eng. Abdikarin** for his support and guidance. As my supervisor, he not only helped me but also shared important ideas that improved this project. His advice and encouragement made it possible for me to complete this research. I also want to thank the **faculty of Computing at SIMAD University** for their technical support and for allowing me to use the laboratory facilities. A special thanks to my teammates, especially **Alisaciid aadan xasan,** for their teamwork and dedication, which made this journey easier and more enjoyable. I am also thankful to my family and friends for their love, support, and encouragement. They gave me the strength and motivation to finish my work. Finally, thank you.

# ABSTRACT

Liver disease poses a significant global health burden, making early detection and

accurate prediction crucial for effective management. This study presents a machine

learning-based approach to predicting liver disease, leveraging various algorithms to

identify the most effective predictive model. The dataset employed includes diverse

biochemical markers and demographic features such as age, gender, total bilirubin, direct

bilirubin, alkaline phosphatase, alanine aminotransferase, aspartate aminotransferase,

total proteins, albumin, and the albumin/globulin ratio. We explored the predictive

capabilities of several machine learning models, including random forest, logistic

regression, decision tree, gradient boosting, and neural networks. Each model's

performance was evaluated based on metrics. Among the models tested, the gradient

boosting model emerged as the most effective, achieving an impressive training accuracy

of 97.65% and a testing accuracy of 97.37%. This model demonstrated superior

performance in terms of overall accuracy and robustness compared to other models.

**Keywords:** Liver disease prediction, Machine learning, Random- forest, Logistic

regression.

Table of Contents

[DECLARATION II](#_Toc204863811)

[DECLARATION. B III](#_Toc204863812)

[ACKNOWLEDGEMENTS IV](#_Toc204863813)

[ABSTRACT V](#_Toc204863814)

[CHAPTER 1 1](#_Toc204863815)

[INTRODUCTION 1](#_Toc204863816)

[1.1Background 1](#_Toc204863817)

[1.2Problem Statement 13](#_Toc204863818)

[1.3 Research question 14](#_Toc204863819)

[1.4 Benefit of the Study 14](#_Toc204863820)

[1.5 Research objective 14](#_Toc204863821)

[1.6 Significance of the study 15](#_Toc204863822)

[1.7 scope and limitation of study 15](#_Toc204863823)

[1.8Organization of the study for liver disease prediction using machine learning 15](#_Toc204863824)

[CHAPTER 2 17](#_Toc204863825)

[RELATED WORKS 17](#_Toc204863826)

[2.1 Introduction 17](#_Toc204863827)

[2.2 Related Work 17](#_Toc204863828)

[2.3. LITERATURE REVIEW 28](#_Toc204863829)

[CHAPTER 3 30](#_Toc204863830)

[RESEARCH METHODOLOGY 30](#_Toc204863831)

[3.1 Introduction 30](#_Toc204863832)

[3.2 System Architecture 30](#_Toc204863833)

[3.2.1 System Features 31](#_Toc204863834)

[3.2.2 Methodology 31](#_Toc204863835)

[3.2.3Data Collection 31](#_Toc204863836)

[3.2.4 Data Preparation 32](#_Toc204863837)

[3.2.5 Model Selection 33](#_Toc204863838)

[3.3 Model Training 37](#_Toc204863839)

[3.3.1 Prediction Generation 37](#_Toc204863840)

[3.3.2 System Requirements 38](#_Toc204863841)

[3.3.3 Hardware Requirements 38](#_Toc204863842)

[3.6.2 Software Requirements 38](#_Toc204863843)

[CHAPTER 4 40](#_Toc204863844)

[RESULTS AND DISCUSSION 40](#_Toc204863845)

[4.1 Introduction 40](#_Toc204863846)

[4.2 Discussion 40](#_Toc204863847)

[4.3 Overview of the implementation environment 41](#_Toc204863848)

[4.4 Data Preprocessing 41](#_Toc204863849)

[MODEL SELECTION 45](#_Toc204863850)

[4.5 Snapshots of the system 46](#_Toc204863851)

[4.5.1 Front-end 46](#_Toc204863852)

[4.5.2 Back-end 50](#_Toc204863853)

[CHAPTER 5 54](#_Toc204863854)

[CONCLUSION AND LESSONS LEARNED 54](#_Toc204863855)

[5.1 Introduction 54](#_Toc204863856)

[5.2 Conclusion 54](#_Toc204863857)

[5.3 Lessons Learned 55](#_Toc204863858)

[5.4 Future Work 57](#_Toc204863859)

[REFERENCE 59](#_Toc204863860)

LIST OF TABLES

[Table 3. 1 Hardware Requirements 37](#_Toc205027461)

[Table 3. 2: Software Requirement’s 38](#_Toc205027462)

**LIST OF FIGURES**

[Figure 3. 1 This diagram shows the system architecture of liver disease prediction using machine learning. 29](#_Toc205028294)

[Figure 4. 2 mission values of null value 41](#_Toc205028295)

[Figure 4. 3 mission values of duplicate 41](#_Toc205028296)

[Figure 4. 4: Random Forest for Classification Training testing accuracy 43](#_Toc205028297)

[Figure 4. 5: Decision trees model for Classification Training testing accuracy 43](#_Toc205028298)

[Figure 4. 6 Model selection 44](#_Toc205028299)

[Figure 4. 7: Login page 46](#_Toc205028300)

[Figure 4. 8: Signup Page 47](#_Toc205028301)

[Figure 4. 9: Home Page 48](#_Toc205028302)

[Figure 4. 10: import necessary 50](#_Toc205028303)

[Figure 4. 11: Gender Visulization 51](#_Toc205028304)

[Figure 4. 12: Result Visulization 51](#_Toc205028305)

[Figure 4. 13: Age Distribution 52](#_Toc205028306)

# CHAPTER 1

# INTRODUCTION

The liver is one of the most vital organs in the human body, as it plays significant roles in several physiological processes, which include the breakdown of red blood cells, enzyme production, removing damaged or worn-out tissues, and waste product processing. With a growing trend of sedentary lifestyles, a lack of physical activities, increased alcohol consumption, and smoking, diseases related to the liver have become a common encounter nowadays. Liver disease encompasses a range of conditions that affect the liver, including hepatitis, cirrhosis, and fatty liver disease, and it is the cause of millions of deaths every year. The World Health Organization (WHO) report on disease trends categorized liver cirrhosis as the 9th leading cause of death in lower- and middle-income countries as of 2022.

# 1.1Background

Liver is one of the largest organs that is present in the upper right part of abdominal cavity, and it is also the second largest organ after skin. It is wedge shape. And it is also the largest gland of the body which secretes chemical substances called hormones. Liver performs more than 500 functions in human body and also supports most of the organ which is vital for our survival. In adults, it is observed that the liver weighs about 2% of body weight, in Males the liver weighs about 1.4 – 1.8 KGS, in females the liver weighs about 1.2 – 1.4 KGS and in new born it weighs 150g. Liver disease is the swelling of the liver caused by toxic substances, bacteria or inherited disease which causes the liver not to function properly as it is essential for digestion and get rid of bacteria. Liver diseases are commonly found in people around the age group of 40-60 years and it is found mostly in men. There is around 10 lakh people diagnosed with liver disease every year and a total of 1.4 lakh deaths in a year in India. Machine Learning is a process which is used to discover patterns in huge data/ large data set to enable decision, thereby allowing machines to go through a learning process (i.e. supervised, unsupervised and semi-supervised or reinforced) .(Gupta et al., 2022).

Machine Learning is a part of Artificial intelligence AI which simulates human intelligence into machines that is used to program to think like humans and mimic their actions. In other words, ML helps the system to gain knowledge without any specific knowledge. In Supervised algorithm, the user inputs and the outputs are used for training process and accuracy prediction.The process of limiting the number of input variables in order for the machine learning algorithm to train the model more-faster is known as feature selection. It reduces the computational complexity and makes it easier to interpret. Machine learning has extended its space to health care as well. One of the problems faced by health care the increase in the number of patients. Applications of Machine Learning can be potentially boosting accuracy for treatment accuracy. The presence of enzymes in the blood can be used to identify liver disease. In this paper we are using liver patient dataset to predict whether the patient is having liver disease or not.(Gupta et al., 2022)

Liver disease is the most dangerous disease and the major reason behind mortality among human beings. The liver is among the largest and most essential organs of our body. Being one of the exocrine gland’s liver secrets bile into the intestine. Observing the last two decades, liver disease is among most top 12 dangerous causes of death and is among the top reason for death cause in adults who are in the age group from 45-54. Fatty liver disease is caused due to extra deposit of fat in the liver cells. Steatosis is the process in which there is the deposition of fat in our liver cells and steatosis can be caused by consumption of alcohol, contaminated food, pickles, etc., and also by metabolic syndrome. Non-cholestatic cirrhosis and hepatocellular carcinoma are majorly caused by fatty liver diseases(V. Singh et al., 2021).

In the human body, liver is considered as the main organ, which plays a central role in several bodily functions. In the human body the production of glucose, processing waste products, producing protein, removing worn-out tissue or cell, blood clotting to cholesterol, and iron metabolism are the core functions of the liver, Somehow, Liver disease cased due to the failure of any of these functions. According to the World Gastroenterology Organization (WGO )and World Health Organization (WHO), 35 million death cases causes occur due to liver failure.Liver diseases are basically classified into two classes that are acute and chronic. The acute liver disorder is an uncommon failure where fast debilitating of liver capacity results in coagulopathy, habitually with an International Normalized Ratio (INR) of greater than 1.5, and variation in the intellectual status (encephalopathy) of an earlier healthy person. For the most part, the youngsters are influenced because of acute liver disorder which conveys a high proportion of death cases(Khan et al., 2019).

The chronic liver disorder is a disease process of the liver which includes a procedure of dynamic devastation and recovery of the liver parenchyma prompting fibrosis and cirrhosis. Substantial Alcohol Consumption is the most widely recognized reason for liver disease. While drinking alcohol, the liver continues its role from the normal to focusing mostly on renovating alcohol into fewer toxic forms. Obesity: People with the substantial fat on their muscles mostly accrue around the liver, cause fatty liver disease. Diabetes: Diabetes patients have the risk of 50 percent liver disease, due to the high level of insulin that results in fatty liver disease. Nowadays data mining and machine learning become basic in healthcare due to its strategies e.g. for example classification, clustering, association rule mining for discovering frequent patterns pragmatic for disease prediction on medical data. This study also includes the comparison of previously used models that are based on MLP, KNN, SVM, J48, RF, DS, RT and LR. The performance of each technique on the dataset is taken from UCI Machine Learning Repository is evaluated using MAE, RAE and Accuracy metrics.(Khan et al., 2019).

Globally, liver disease has become an alarming and life-threatening issue. Machine learning algorithms can early help in early diagnosis to reduce risk. Analyzing the previous studies showed low performance. Hence this research aims to achieve more satisfactory performance. This research aims to determine the accuracy of several popular machine algorithms—XGBoost, SVM, logistic regression, random forest, AdaBoost, Decision Tree, and K-NN—to predict liver diseases by analysing different data sets and comparing their performances. The main contributions are to find out the correlation matrix with the outcome, the model performance of the lowest split to the higher split of the training set, and the best split of the training set. the liver assumes a central part in the coagulation cycle and the upkeep of hemostasis. It is considered one of the significant organs in the human body and the major vital organ with internal functions, such as creating compounds, handling waste items, and eliminating destroyed tissues or cells(Ghosh et al., 2021)**.**

The liver presents with different illnesses that inevitably can cause it harm and other sorts of related illnesses. Approximately 35 million individuals died because of chronic infections, as indicated by the World Gastroenterology Organization (WGO) and the World Health Organization (WHO), and liver disorder is one of the illnesses mentioned. Because of high alcohol consumption, breathing of harmful gases, and ingesting spoiled food, pickles, and medications, patients with liver sickness have persistently worsened. An early analysis of liver issues may raise a patient’s recovery rate. Liver ailments can be analyzed by examining the enzyme levels in the blood . Presently, the prevalence of alcoholic liver disease (ALD) has been growing. In some studies, age-standardized mortality associated with HCV is shown to have risen from 7.17 per 100,000 people in 2007 to 8.14 per 100,000 people in 2013, accompanied by a stamped decline when patients started direct-acting antiviral therapy treatments. Here, HCV represents the hepatitis C infection. Even though HCV contamination has improved, the appearance of ALD has increased, specifically in non-Hispanic Asians. There was a higher death rate for hepatitis B infection contamination among non-Hispanic Asians(Ghosh et al., 2021)**.**

The liver is a large, meaty organ that sits on the right side of the belly. Weighing about 3 pounds, the liver is reddish-brown in colour and feels rubbery to the feel. The liver has two large sections, called the right and the left lobes. The gallbladder sits below the liver, along with parts of the pancreas and intestines. The liver and these organs behaviour together to digest, absorb, and process food. 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. Liver disease is any trouble of liver function that causes sickness. The liver is responsible for many dangerous functions within the body and should it become diseased or damaged, the loss of those functions can cause significant injury to the body(Priya et al., 2018).

Liver disease is also referred to as hepatic disease. Liver disease is a large term that covers all the potential problems that cause the liver to fail to perform its designated functions. Usually, more than 75% or three quarters of liver tissue needs to be affected before a decrease in function occurs. Several diseases states can disturb the liver. Some of the diseases are Wilson's disease, hepatitis, 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. Fatty liver is a revocable condition where large vacuoles of triglyceride fat acquire in liver cells via the process of limit. It can occur in people with a high level of alcohol consumption as well as in people who never had alcohol.Hepatitis (usually caused by a virus spread by excess contamination or direct contact with infected body fluids). Cirrhosis of the liver is one of the most serious liver diseases. It is an action used to indicate all forms of diseases of the liver characterized by the significant loss of cells(Priya et al., 2018).

The liver gradually contracts in size and becomes leathery and hard. The regenerative action continues under liver cirrhosis but the progressive loss of liver cells exceeds cell replacement. The risk of liver cancer is higher in those who have cirrhosis or who had valid types of viral hepatitis; but more often, the liver is the site of secondary (metastatic) cancers spread from other organs. PSO feature extraction model for liver dataset and applied an improve probability in many medical applications, such as training artificial neural networks, linear constrained function optimization, wireless network optimization, data classification, and many other areas where GA can be applied. Computation in PSO is based on a swarm of processing elements called particles in which each particle represent a candidate solution. Random forests are machine learning regression method for classification that drive by constructing liver data into a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. It is unexcelled in accuracy among current algorithms. It output classification efficiently on large liver dataset. It can handle thousands of input attributers without variable deletion. It gives estimates of what variables are important in the classification. Random Forests grows many classification trees. To classify a new liver object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and says the tree "votes" for that class. The forest chooses the classification having the most votes. (Priya et al., 2018).

Although, liver is the largest organ of the body and also it is essential for digesting food and releasing the toxic element of the body. The viruses and alcohol use lead the liver towards liver damage and lead a human to a life-threatening condition. There are many types of liver diseases whereas hepatitis, cirrhosis, liver tumours, liver cancer, and many more. Among them liver diseases and cirrhosis as the main cause of death. Therefore, liver disease is one of the major health problems in the world. Every year, around 2 million people died worldwide because of liver disease. According to the Global Burden of Disease (GBD) project, published in BMC Medicine, one million peoples are died in 2010 because of cirrhosis and million are suffering from liver cancer(Rahman et al., 2019).

Machine learning has made a significant impact on the biomedical field for liver disease prediction and diagnosis. Machine learning offers a guarantee for improving the detection and prediction of disease that has been made an interest in the biomedical field and they also increase the objectivity of the decision-making process. By using machine learning techniques medical problems can be easily solved and the cost of diagnosis will be reduced. Therefore, we used different classification techniques for the classification of patients have liver disease or not. Six machine learning techniques have been applied including LR, KNN, DT, SVM, NB, RF and the performance of these techniques were estimated on various perspectives such as accuracy, precision, recall, f-1 score. Moreover, the performance was compared using the receiver operative characteristic(Rahman et al., 2019).

The Liver is an intense and vital organ in the human body. It consists of two large portions, known as the left projections and privilege. The Gallbladder, Pancreas, and Stomach are some organs that reside nearby the Liver. Liver and their organs collaborate to ingest and process substances.A job of the Liver is to remove the harmful substances from the blood beginning from the framework, which is related to the stomach prior crossing it to what-so-ever is left in the body. Liver disease can be adopted either from outside factors like infections or alcohol. Corpulence has additionally been related to this illness. Overtime harm to the Liver can cause liver failure and sometimes lead to a life-threatening condition. The liver is located in the upper part of the gastrointestinal tract of the human body, and its weight in men ranges between 1400–1800 g and in women between 1200–1400 g. It performs important functions related to digestion, metabolism, releasing toxins, immunization and nutrient storage. That is why some liver diseases can even lead to death.liver diseases are categorized based on their aetiology and effect on the liver(Dritsas &Trigka, 2023).

The aetiology may include infection, injury, exposure to drugs or toxic substances, a process, or a genetic abnormality (such . The above causes can lead to hepatitis, cirrhosis, and stones that can increase in size and cause blockages, fatty infiltration and, in rare cases, liver cancer. Genetic abnormalities can also interfere with vital functions of the liver and lead to the deposition and concentration of harmful components, such as iron or copper. Thus, the liver is permanently injured and cannot function properly. The main causes of cirrhosis of the liver include alcoholism, non-alcoholic fatty liver disease, chronic hepatitis C, and chronic hepatitis B. Non-alcoholic fatty liver disease (NAFLD) is one of the main liver diseases and is characterized by an accumulation of lipids in the liver. If there is inflammation and injury to the liver cells, it is called “non-alcoholic steatohepatitis” . Cirrhosis is also one of the most serious liver diseases(Dritsas &Trigka, 2023).

This disease causes healthy tissue to be replaced by scar tissue. There are two main forms of hepatitis, acute hepatitis, where the liver becomes inflamed at a rapid rate, and chronic hepatitis, where the liver becomes inflamed and destroyed slowly over a long period of time. Although hepatitis can be caused by any of the above, it is most often caused by infection with a virus in a group called hepatitis viruses. These viruses have been named, in the order they were discovered, as hepatitis A, B, C, D, and E viruses. liver disease is usually associated with alcoholism or hepatitis, but obesity and diabetes are becoming a growing threat of potentially fatal liver damage. Advanced fatty liver disease increases a person’s risk of death by nearly sevenfold. It is a silent “killer” and if the symptoms associated with fatty liver damage appear, the situation is already difficult(Dritsas &Trigka, 2023).

therefore, accurate and efficient diagnosis of a person at initial traceable and risk of Liver Diseases might provide huge aids for prevention, diagnosis, or the right treatment. From the past decade, biopsy was the method used to stratify the patients, called a diagnostic reference standard for finding of liver with fatty intrusions. This procedure is large expensive & cumbersome; it might also provide the sampling errors and side effects when the application of this method conducted. Hence, traditional detection techniques now do not prevail of much importance until a late-stage diagnosis. So, apart from clinical methods, modern-age technology believes in highly sophisticated computational techniques which help to simplify work extensively with a higher rate of accuracy and minimal error. Artificial Intelligence and machine learning are among the major domains and applications of the computer science. It uses the implementation of the computer algorithms to determine the massive dataset patterns and also helps to forecast the different data-based outcomes. In the presence of the clinical data, ML plays a vital role for the decision making in the medical field. Various machine learning models of development will act as essential aids to determine disease and would result in better decisions. It allows the hospital resources optimization by dividing the right patients with critical various earlier risk factors(Choudhary et al., 2021).

To detect disease, healthcare professionals need to collect samples from patients which can cost both time and money. Often, more than one kind of test or many samples are needed from the patient to accumulate all the necessary information for a better diagnosis. The most routine tests are urinalysis, complete blood count (CBC), and comprehensive metabolic panel (CMP). These tests are generally less expensive and can still be very informative. The liver has many functions such as glucose synthesis and storage, detoxification, production of digestive enzymes, erythrocyte regulation, protein synthesis, and various other features of metabolism. Chronic liver diseases include chronic hepatitis, fibrosis, and cirrhosis. (Mostafa et al., 2021).

Hepatitis can occur from viral infection (e.g., hepatitis c virus) or auto-immune origin. Inflammation from hepatitis infection can cause tissue damage and scarring to occur in the liver. Moderate scarring is classified as fibrosis, while severe liver damage/scarring is classified as cirrhosis. Fibrosis and cirrhosis can also occur from alcoholism and non-alcoholic fatty liver disease. When liver disease is diagnosed at an earlier stage, in between infection and fibrosis but before cirrhosis, liver failure can be avoided. Tests, such as a CMP and biopsy, can be conducted to diagnose all forms of liver disease. A CMP with a liver function panel can detect albumin ALB, alkaline phosphatase ALP, alanine amino-transferase ALT, aspartate amino-transferase AST, gamma glutamyl-transferase GGT,creatine CREA, total protein PROT, and bilirubin BIL .Diagnosis of a certain liver disease and discovery of its origin are made by interpreting the patterns and ratios of circulating liver-associated molecules measured with the CMP test and compared to values normalized with a patient’s age, sex, and BMI. Aminotransferases, AST, and ALT are enzymes that participate in gluconeogenesis by catalyzing the reaction of transferring alphaamino groups to ketoglutaric acid groups. AST is found in many tissue types and is not as specific to the liver but may denote secondary non-hepatic causes of liver malfunction. ALT is found in high concentrations in the cytosol of liver cells(Mostafa et al., 2021).

Liver cell injury can cause the release of both aminotransferases into circulation.Liver-related disease accounts for 70% of deaths worldwide. There is a need to find better ways to detect and diagnose liver disease with more accuracy. Most importantly, tests of liver function need to be available and affordable to patients. To avoid the expensive and invasive tests, the application of statistical machine learning techniques to CMP results for the extraction of information for a clinician might be helpful for diagnosis. Exploratory data analysis methods are extremely important in healthcare; they can predict patterns across data sets to facilitate the determination of risk or diagnostic factors for disease with more speed and accuracy. Using machine learning algorithms to predict disease is made possible by increasing access to hidden attributes in medical data sets. Various kinds of data sets, such as blood panels with liver function tests, histologically stained slide images, and the presence of specific molecular markers in blood or tissue samples, have been used to train classifier algorithms to predict liver disease with good accuracy (Mostafa et al., 2021).

In the human body, the liver organ plays a very important role in many functions like decomposition of red blood cells, etc. The liver is the largest body organ and located in the upper right position of our abdomen liver disease refers to any disorder with the liver. This disease contains many conditions like inflammation (hepatitis B, C) from infectious, non-infectious causes (chemical or autoimmune hepatitis),Tumours, malignant and scarring of the liver (Cirrhosis) and Metabolic Disorders. In this paper, six classification algorithms J-48, Random Forest, Logistic Regression, SMO (Support Vector Machine), IBK (k nearest Neighbour), Logistic Regression, Naive Bayes have been considered for implementation and comparing their results based on the ILPD (Indian Liver Patient Dataset)(J. Singh et al., 2020).

Liver is a major organ of the body which is located in the right-side portion and above of the abdominal cavity but just beneath the diaphragm, top of the stomach, under the rib cage. The size of a liver of the human being is large and reddish-brown in colour. Liver is a most important organ of the body of every life form who are vertebrates. Liver should do their proper functions to maintain the body healthy and feet. Liver has major functions like, digestion of foods which we uptake from day to night, secretion of many types enzymes which are required to carry out different biochemical pathways such as, glycolysis, gluconeogenesis, tri-carboxylic acid cycle, protein synthesis, beta-oxidation of fatty acids, detoxification of xenobiotic compounds etc(Veeranki& Varshney, 2022).

Liver is a very much essential compartment of the body as it helps in foods digestion resulting in converting them into simple nutrients from complex compounds by the help of various enzymes present within the body. Major as well as important biochemical pathways are accomplished by the enzymes present in the liver cells. Serum glutamic pyruvic transaminase SGPT and serum glutamic-oxaloacetic transaminase SGOT are enzymes present in liver cells in normal health condition but when these enzymes are released into the bloodstream as well as increase their amount in the blood, indicate that the liver does not function correctly and the liver has been damaged. Other names of SGPT and SGOT are alanine transaminase ALT and aspartate aminotransferase AST. A high amount of these two enzymes in the bloodstream indicate that the liver will be damaged completely in future, if it is not cured properly. Liver can be damaged by various viral attacks, fibrosis, cancer, alcohol consumption, bacterial and parasite attack, appropriate metabolic functions, abnormal gene function etc. These phenomena can alter liver structure and responsible for liver malfunctioning(Veeranki& Varshney, 2022).

Liver is a central organ of the body so it plays a vital role to maintain a good health. Digestion of foods to production of energy and growth as well as detoxification of the toxic compounds is main functions of the liver. Liver is exposed to the various toxic compounds, various pathogens, xenobiotic compounds etc. Liver produces different enzymes, hormones which are regulated properly and regularly to defend those kinds of threats and makes a sound health. The biological chemistry is a system of the body which should be regulated by the biochemical factors itself and build up a sound health to live the life healthier and happier(Veeranki& Varshney, 2022).

The liver is the largest internal organ in our human body. The liver has two lobes, left lobe and right lobe. The liver weight is approximately 3 pounds, it’s a reddish-brown color. The gallbladder is located under the liver. The main important role of the liver is to remove the toxic and harmful substances from the blood before distribution to different parts of our body. Liver disease is also considered one of the most dangerous and deadliest diseases faces in the globe. it is very difficult to identify in early stages of liver disease even liver tissue has damaged moderately, in this case many medical expert systems difficult to identify the disease(Shaheamlung et al., 2020). This leads to fail in treatment and medication. In order to avoid this early prediction is crucial to give proper treatment and save life of patient Machine learning is a branch of Artificial Intelligence, which help the computer to think like human and can take their own decision without human intervention. Due to rapidly development in Artificial Intelligent, Machine learning has lots of advancement in diagnosis of difference types of disease(Shaheamlung et al., 2020). More-over Machine learning algorithm gives us more accurate prediction and performance.Machine learning (ML) is defined as a discipline of artificial intelligence (AI) that provides machines the ability to automatically learn from data and past experiences to identify patterns and make predictions with minimal human intervention. It has different types that solve each of our problems and one of them is supervised machine learning, unsupervised machine learning and reinforcement learning. It also has a lot of algorithms that allow us to make predictions about liver disease. It also has models that are trained on our data. and to test what we are allowed to do are the models that have machine learning(Shaheamlung et al., 2020).

Machine learning has been broadly divided into different types. SUPERVISED LEARNING: In easy word, supervised learning is types of learning method with the help of supervisor, teacher or instructor. It consists of training set of patterns associated with label data and makes it easy for algorithm from input to output and also easy to learn and predict. Some of supervised learning are classification such as KNN, SVM, Naïve Bayes, Neural network regression as linear and polynomial, Decision tree and RandomForest. Developed prediction based on both input and output data. UNSUPERVISED LEARNING: Unsupervised learning is also known as clustering. In unsupervised learning there is no training data set, no label and unknown output data. This type of learning method is like self-guide learning method. Some of the supervised learning methods are clustering such as K-Means clustering, SVD and PCA. REINFORCEMENT LEARNING:This is a type of machine learning based on agent, action, state, reward and environment(Shaheamlung et al., 2020).

Liver is an essential organ of our body. There is a great need for an early detection of liver disease so as to prevent complete liver failure, which can result in patient’s death. For the proper diagnosis, it is necessary to evaluate some of the main attributes of liver patient’s dataset . Some of the main attributes of liver disease include, Total-bilirubin, direct-bilirubin, alkaline-phosphotas, total-protein, albumin and globulin-ratio. It is the challenging task for doctors to accurately predict the liver disease. Various classification techniques are used to classify the data and predict the liver disease through the datasets of liver patients, having access to classification algorithms with large amount of data will help clinicians make better decisions and ultimately improve patient outcomes with an accurate prediction of liver disease.‖ This paper shows a survey about the classification techniques that can be used for the prediction of liver disease and gives an idea for future work, that which classification technique can be utilised further for diagnosis of the liver disease(Khan et al., 2021).

The diagnosis of liver disease or condition depends on the information that includes risk factors that make accurate diagnosis difficult. These risk factors include resource and organizational limits, conflicts, ambiguity, and uncertainty. Many symptoms are vague and vary from person to person. Several diagnostic tests are costly, seldom performed, and often do not provide a black-and-white result. In addition, cognitive bias and improper use of heuristics are common occurrences during the diagnostic phase among physicians.(Ding et al., 2022).

The liver is the primary organ of the body, responsible for metabolism and protein absorption. Many diseases affect the liver, including hepatitis, fibrosis, and cirrhosis. Liver cirrhosis and liver cancer stand at 11th and 16th rank as the most common causes of death globally, with yearly more than 1 million and 0.7 million deaths(Patnaik et al., 2022).

Liver disorder is a life-threatening disorder with high mortality rate every year. Prognosis of the liver disorder includes various steps starting with the routine urine and blood tests. Based on the symptom’s observed patient is advised for LFT. Several surveys have been performed to effectively diagnose the liver disorder using various machine learning algorithms. the derivation of concealed information from massive data stores is performed by potential mechanism called Data Mining. The data mining strength in healthcare domain implements the precise use of data and analytics there by classifying the inefficiencies and analyse the best techniques to augment care and reduce the cost. The complexity of healthcare system necessitates a proper technology adoption for mining the huge hidden information . Cancer contributes to the maximum death rate and cancerous liver being most prominent amongst them. Every year people suffering from cancer all over the world exceed 700,000. Liver cancer contributes to over 600,000 deaths every year. The symptoms that are diagnosed in clinical trials are used to predict the disease. The patient is advised to undergo LFT if predicted with some liver disorder. (Gogi & Vijayalakshmi, 2018).

The liver is the second largest internal organ in the human body, playing a major role in metabolism and serving several vital functions, e.g. Decomposition of red blood cells, etc. Its weight comes around three pounds. The liver performs many essential functions related to digestion, metabolism, immunity, and the storage of nutrients within the body. These functions make the liver as an important organ, without this, body tissues would quickly die from lack of energy and nutrients. There are number of factors which increase the risk of liver disease.(Vijayarani & Dhayanand, 2015)

The liver plays an important role in many bodily functions from protein production and blood clotting to cholesterol, glucose and iron metabolism. It has a range of functions, including removing toxins from the body, and is crucial to survival. The loss of those functions can cause significant damage to the body. When liver is infected with a virus, injured by chemicals, or under attack from own immune system, the basic danger is the same – that liver will become so damaged that it can no longer work to keep a person alive. Liver disease caused by hepatotrophic viruses

# 1.2Problem Statement

Globally, liver disease is a significant health concern worldwide, with millions of people affected by various liver disorders. Early detection and accurate prediction of liver disease are crucial for effective treatment and prevention of further complications.

There are many challenges that is the limited availability of high-quality datasets containing comprehensive patient information and representative samples. Also, an existing models may lack the required accuracy and reliability, leading to missed diagnoses, suboptimal treatment plans, and increased healthcare costs. Additionally, existing models may lack the required accuracy and reliability, leading to missed diagnoses, suboptimal treatment plans, and increased healthcare costs.these challenges and developing an accurate liver disease prediction model would significantly improve early diagnosis, provide timely interventions, and ultimately enhance patient outcomes.

# 1.3 Research question

1. How can we develop accurate and reliable machine learning models for predicting liver diseases?
2. How can we develop predictive models that integrated to clinical practice to support healthcare providers in decision-making processes?
3. How can we develop web or application that allow us doctors or users enter data to predict the data that they entered?

# 1.4 Benefit of the Study

The project holds great promise for the improved management of liver disease. The proposed system can serve as a means for early prediction and diagnosis aids, allowing healthcare professionals to make sound decisions, optimize resources, and improve patient outcomes.

The benefits of the project are

More Effective and quicker Diagnosis: The diagnostic predictions will be fast and accurate, decreasing the time taken for diagnosis.

Lower cost: It may reduce the use of costly and invasive diagnostic techniques, and thus the cost of care.

# 1.5 Research objective

1. To develop accurate and reliable machine learning models for predicting liver diseases.
2. To develop web or application that allow us doctors or users enter data to predict the data that they entered.
3. To integrate predictive models into healthcare systems to support early detection and personalized treatment of liver diseases.

# 1.6 Significance of the study

The advantage of diagnosing liver disease using machine learning is to have a machine that can help us predict whether a person has liver disease or not and it takes less time to diagnose, that is the big advantage. That we get if we do this title called predicting liver disease using a machine. A quick response is given to patients with liver disease. A doctor uses a liver disease predictor using a machine to determine whether a person has liver disease. Patients with liver disease receiving rapid response.

# 1.7 scope and limitation of study

The Study Will Concentrate on Liver Disease Prediction, headquarter in Mogadisho Somalia,

In Content, Risk Assessment and Prediction: The validated models are utilized to assess an individual's risk of developing liver disease based on their specific characteristics and test results. This risk assessment can provide insights into the likelihood and severity of liver disease progression.

Clinical Application: The developed predictive models can be integrated into clinical practice to support early detection, intervention, and personalized management of liver disease. Healthcare providers can use these models as decision-support tools to identify high-risk individuals who may benefit from targeted interventions, such as lifestyle modifications, medication, or surveillance.

Liver disease prediction in research holds great potential for improving patient outcomes by enabling early detection and intervention. By identifying individuals at high risk, healthcare resources can be allocated more efficiently, and preventive measures can be implemented to mitigate the progression of liver diseases. And the Study carried out Jan to Jun 2024

# 1.8Organization of the study for liver disease prediction using machine learning

Chapter 1: Introduction

This chapter serves as an introduction to the study, outlining the research plan. It includes the background of the liver disease prediction and its significance,

Chapter 2: Literature Review

This chapter discusses existing research on predicting liver disease with the help of machine learning.

Chapter 3: Methodology

This chapter describes the technique utilized in this study, i.e., predicting liver disease using machine learning.

Chapter 4: Implementation and Coding and Results

This chapter addresses the implementation of machine learning models for the prediction of liver disease using software tools. It also explains the difficulties encountered during implementation.

Chapter 5: Conclusion and Future Work

This chapter wraps up the study on liver disease prediction by presenting the research findings and offering recommendations for future research in this field.

# CHAPTER 2

# RELATED WORKS

# 2.1 Introduction

Machine learning (ML) has emerged as a powerful tool in the field of medical diagnostics and prognostics, including predictive models for liver diseases. This literature review aims to summarize recent research that applies ML algorithms to improve the accuracy and efficiency of liver disease diagnosis and risk assessment.

# 2.2 Related Work

Dhamodharan et.al(2014) , has predicted three major liver diseases such as Liver cancer, Cirrhosis and Hepatitis with the help of distinct symptoms. They used Naïve Bayes and FT Tree algorithms for disease prediction. Comparison of these two algorithms has been done based on their classification accuracy measure. From the experimental results they concluded the Naïve bayes as the better algorithm which predicted diseases with maximum classification accuracy than the other algorithm.

Rosalina et al (2010) predicted a hepatitis prognosis disease using Support Vector machine (SVM) and Wrapper Method. Before classification process they used wrapper methods to remove the noise features. Firstly SVM carried out feature selection to get better accuracy. Features selection were implemented to minimize noisy or irrelevance data. From the experimental results they observed the increased accuracy rate in the clinical lab test cost with minimum execution time. They have achieved the target by combining Wrappers Method and SVM techniques.

Omar S. Soliman et al (2014) has proposed a hybrid classification system for HCV diagnosis, using Modified Particle Swarm Optimization algorithm and Least Squares Support Vector Machine (LS-SVM). Feature vectors are extracted using Principle Component Analysis algorithm. As LS-SVM algorithm is sensitive to the changes of values of its parameters, Modified-PSO Algorithm was used to search for the optimal values of LS-SVM parameters in less number of iterations. The proposed system was implemented and evaluated on the benchmark HCV data set from UCI repository of machine learning databases. It was compared with another classification system, which utilized PCA and LS-SVM. From the experimental results the proposed system obtained maximum classification accuracy than the other systems.

Karthik et.al (2011) were applied a soft computing technique for intelligent diagnosis of liver disease. They have implemented classification and its type detection in three phases. In the first phase, they classified liver disease using Artificial Neural Network (ANN) classification algorithm. In the second phase, they generated the classification rules with rough set rule induction using Learn by Example (LEM) algorithm. In the third phase fuzzy rules were applied to identify the types of the liver disease.

Chaitrali S. Dangare et.al (2012)has analyzed prediction systems for Heart disease using more number of input attributes. The data mining classification techniques, namely Decision Trees, Naive Bayes, and Neural Networks are analyzed on Heart disease database. The performances of these techniques are compared, based on accuracy. Authors’ analysis shows that out of these three classification models Neural Networks has predicted the heart disease with highest accuracy.

In an experiment Gunasundari (2013) found conversional image processing operations, neural networks and Genetic algorithm gives successful result for liver disease disorder diagnosis. In future liver disease disorder diagnosis extended in many directions. Such as using effective algorithms and more texture feature technique algorithms. CART uses a purity-based measure, and the algorithm splits the training data set based on how probably the subsets become purer for a class, and it spends more time to generate smaller trees. CART and C4.5 both algorithms are gives good result with oversampling for liver disease disorder dataset. These algorithms could reduce the minor class increment to smaller percentage. Decision tree algorithm does not give high priority for minor classes for that reason using duplication in BUPA liver disease disorder dataset, increase the number of instances of minor class and proceed with two decision tree algorithms and both algorithms gives good result in insufficiency of liver disease disorder data. Case Based Reasoning (CBR) and Classification and Regression Tree (CART) techniques could be useful to detect the liver disease [8]. Feature selection plays a vital role in text categorization. A range of different methods have been developed, each having unique properties and selecting different features. We show some results of an extensive study of feature selection approaches using a wide range of combination methods

Bendi (2012), proposed a Modified Rotation Forest algorithm to calculate the accuracy of the liver classification techniques in UCI liver dataset using the combo of feature selection technique and selected classification technique algorithm. Over the past few years, the increasing attention on severe challenges in medical diagnosis process such as sharply increased elderly patients, limited medical personnel, has led to a number of contributions in the areas of the intelligent medical diagnosis methods. The early contributions can be found on the neural networks, it provides a new significant way for intelligent medical diagnosis.

A model proposed by Kiruba(2014) on intelligent agent based system to hike a precise and accurate of diagnosis system. C4.5 decision tree algorithm and Random tree algorithm are used to predict. Two different types of liver disease disorder dataset are combined and predict the accuracy of the disease. And then conclude these both algorithms gives very good accuracy for diagnosing liver disease disorder. Liver abscess is the commonest cause of hepatomegaly and it is due to amoebiasis, followed by fatty liver, congestive cardic failure, hepatocellular carcinoma, and viral hepatitis seen only in few patients

P.Rajeswari,G.Sophia Reena et al.,(2010)has proposed the data classification is based on liver disorder. The training dataset is developed by collecting data from UCI repository consists of 345 instances with 7 different attributes. This paper deals with results in the field of data classification obtained with Naïve Bayes algorithms. FT tree algorithms, and KStar algorithms and on the whole performance made know FT Tree algorithm when tested on liver disease datasets, time taken to run the data for result is fast when compare to other algorithm with accuracy of 97.10%Based on the experimental results the classification accuracy is found to be better using FT Tree algorithm compare to other algorithms .

Sa’diyah Noor Novita Alfisahrin, Teddy Mantoro et al., (2013) have proposed to identify if the patients have the liver disease based on the 10 important attributes of liver disease using a Decision Tree, Naive Bayes , and NB Tree algorithms. The result shows NB Tree algorithm has the highest accuracy; however the Naïve Bayes algorithm gives the fastest computation time.For future study, the performance of NB Tree algorithm will be the target of improvement of the accuracy by finding the most significant factor in identifying liver disease patients.For future study, the performance of NB Tree algorithm will be the target of improvement of the accuracy by finding the most significant factor in identifying liver disease patients .

S.Dhamodharan (2014) has proposed there are many liver disorders that require clinical care of the physician. They predict three major liver diseases such as liver cancer, cirrhosis, hepatitis with the help of distinct symptoms. The primary goal is to predict the class types from classes such as liver cancer, cirrhosis, hepatitis and ǲno diseasesǳ. In this paper Naïve Bayes and FT tree algorithm accuracy are compared and the result is obtained. The result concludes that the accuracy of Naïve Bayes algorithm is much better than the other algorithms.

S. E. Seker, Y. Unal, Z. Erdem, and H. Erdinc Kocer et al.,2014 has proposed have applied the data mining techniques, such as KNN, SVM, MLP or decision trees over a unique dataset, which is collected from 16,380 analysis results for a year. This study can be useful for further studies like reducing the number of analysis, since the prediction can be correlated and furthermore the correlation can be utilized for detecting the anomaly on the analysis .

A.S.AnneshKumar,Dr.C.JothiVenkateswaran et al.,(2015)has proposed describes the categorization of liver disorder through feature selection and fuzzy K-means classification. Various liver disorders also share same attribute values and it needs more effort to classify liver disorder type correctly with basic attributes. So Fuzzy based classification gives better performance in these confusing classes and achieved above 94 percentage accuracy for each type of liver disorder .

P.Thangarajul, R.Mehala et al., (2015) has proposed to analyze the data of liver diseases using particle swarm optimization algorithm (PSO) with K Star classification. In two aspect for classifying the existence of disease are not.The proposed algorithm enhanced the performance of accuracy when compared to existing classification algorithms. PSO-Kstar algorithm is best suitable algorithm for the classification of liver disorders as it improved the performance in prediction accuracy as discussed earlier .PSO-KStar algorithm is considered is one of the good data mining algorithm with respect to understandability, transformability and accuracy gives 100% .

Onwodi Gregory (2015) has proposed two real liver patient datasets were investigated for building classification models in order to predict liver diagnosis. Eleven data mining classification algorithms were applied to the datasets and the performance of all classifiers are compared against each other in terms of accuracy, precision, and recall. Based on the experimental results the classification accuracy is found to be better using FT Tree algorithm compare to other algorithms., it also shows the enhanced performance according to the attributes and it gives 78.0% of Accuracy, 77.5% of Precision, 86.4% of Sensitivity and 38.2% of Specificity results respectively.

Dr.S.Vijayarani, Mr.S.Dhayanand et al., (2015) has proposed description of this research work is to predict liver diseases using classification algorithms. The algorithms used in this work are Naïve Bayes and support vector machine (SVM Comparisons of these algorithms are done and it is based on the performance factors classification accuracy and execution time. From the results, this work concludes the SVM classifier is considered as a best classification algorithm because of its highest classification accuracy values. On the other hand, while comparing the execution time, the Naïve Bayes classifier needs minimum execution time from the implementation results it is observed that the SVM is a better Classifier for predict the liver diseases and comparing the execution time, the Naïve Bayes classifier needs minimum execution time .

Ebenezer Obaloluwa Olaniyi, Khashman Aadnan et al. (2015) has proposed back propagation neural network and radial basis function neural network are designed to diagnose these diseases and also prevent misdiagnosis of the liver disorder patients. The algorithms were compared with the c4.5, CART, Naïve Bayes, Support Vector Machine (SVM) and concluded that the radial basis function neural network is the optimal model because it has a recognition rate of 70% which has proved more accurate and efficient than the other algorithms.

Tapas RanjanBaitharua, Subhendu Kumar Panib et al., (2016)has proposed focuses on the aspect of Medical diagnosis by learning pattern through the collected data of Liver disorder to develop intelligent medical decision support systems to help the physicians. In this paper the use of several classification (J.48, SVM, Random Forest, etc) algorithms to classify these diseases and compare the effectiveness, correction rate among them. In this paper, a comparative analysis of data classification accuracy using Liver disorder data in different scenarios is presented. The predictive performances of popular classifiers are compared quantitatively.By analyzing the results Multilayer perceptron gives the overall best classification result with the accuracy 71.59% than other classifiers.

Anju Gulia, Dr. Rajan Vohra, Praveen Rani et al., (2014) has proposed to implements hybrid model construction and comparative analysis for improving prediction accuracy of liver patients in three phases. In first phase, classification algorithms are applied on the original liver patient datasets collected from UCI repository. In second phase, by the use of feature selection, a subset (data) of liver patient from whole liver patient datasets is obtained which comprises only significant attributes and then applying selected classification algorithms on obtained, significant subset of attributes. SVM algorithm is considered as the better performance algorithm, because it gives higher accuracy in respective to other classification algorithms before applying feature selection. But, Random Forest algorithm is considered as the better performance algorithm after applying feature selection. In third phase, the results of classification algorithms with and without feature selection are compared with each other. The results obtained from our experiments indicate that Random Forest algorithm outperformed all other techniques with the help of feature selection with an accuracy of 71.8696%.

Anu Sebastian et al(2016). Survival Analysis is an extensively used procedure in the field of medical science. The idea of being able to predict the life expectancy of the subject is of immense value and utility to both, the doctors and the patients. There are three preliminary steps that serve as the elementary foundation of any medical treatment paradigm. The diagnosis stage, the classification stage, the assessment stage, the conclusion stage and finally the treatment stage. All these stages are expected to be accurate to the parameters and effective in their measure to distinctly reflect the quantified magnitude and the intensity of the study of the disease in the context. One of the most widely used classification methodologies that have been used for an extensive assessment of liver diseases, particularly cirrhosis is the Child-Pugh classification method. It is understandable from the extensive study of a voluminous set of cases that the life expectancy of different patients, suffering from different intensities and kinds of liver cirrhosis, is different. Fuzzy Logic, for instance, suits the context like a tailor-made technique.

Insha Arshad et al (2018) Liquor is expended in overabundance by a large number of individuals over the world. Liquor utilization is legitimately connected to perilous liver maladies, for example, cirrhosis which may at last lead to death. Early location of liver illness brought about by overutilization of liquor would help in sparring existences of numerous individuals. By distinguishing liver ailment in its beginning time, it very well may be analyzed in time and may prompt full recuperation in certain patients. This paper proposes identification just as to foresee the nearness of liver sickness utilizing information mining calculations. We will settle on a choice tree for the dataset and afterwards the principles will be created. Subsequent to deciding the principles, we will utilize diverse information mining calculations to prepare and test the dataset to distinguish the liver sickness. The information was gathered from UCI storehouse and our preparation dataset was created. It comprises of 7 unique qualities having 345 occurrences. In the dataset, distinctive classes of blood tests are taken into contemplations which are straightforwardly connected to liver illnesses that may emerge because of unnecessary liquor utilization alongside recurrence of liquor utilization. In light of the sort of liver sickness recognized, the forecast might be proposed.

N. Ramkumar,et al,(2017).Malignant growth is the one of the unsafe infection on the planet. Malignant growth spreads in lungs, liver, bosom, bones and so forth. Liver malignancy is the most hazardous and it will proceed with long-lasting. The side effects of a liver malignant growth are Jaundice, loss of weight, yellow shaded pee, spewing, torment in the upper right stomach area, sweats, fever and amplified liver. The liver malignancy which starts in the liver separated from moving from another piece of the body is called as essential liver disease. A disease which spreads all other pieces of the body lastly it achieves liver is called an auxiliary liver malignant growth. The liver is one of the critical pieces of the human. WHO reviews state out of 100,000 individuals, around 30 individuals have experienced liver malignant growth and generally it influences the African and Asian nations prior. These days it turned into a well-known ailment. The most widely recognized sort of a liver malignant growth is called hepatocellular carcinoma, this specific influences male as opposed to female. The liver malignant growth happens for the most part because of the more liquor utilization. Numerous information mining calculations, artificial insight ideas are utilized to anticipate liver disease. The likelihood of anticipating the liver malignant growth is performed utilizing the Bayes hypothesis with the WEKA tool.

Mafazalyaqeen et al(2017). One of the fascinating and vital subjects among scientists in the field of therapeutic and software engineering is diagnosing disease by considering the highlights that have the most effect on acknowledgements. The subject talks about another idea which is called Medical Data Mining (MDM). Undoubtedly, information mining techniques utilize diverse ways, for example, characterization and grouping to arrange maladies and their indications which are useful for diagnosing. This paper presents another technique for liver illness analysis to help specialists and their patients in finding the sickness side effects and decrease quite a while of diagnosing and counteract passings. The proposed strategy will streamline the tenets discharged from Boosted C5.0 grouping technique with the Genetic Algorithm (GA), to expand the determination time and exactness. So as opposed to utilizing a transformative calculation for creating rules, the hereditary calculation is utilized for improving and diminishing tenets of another calculation. We demonstrate that our proposed methodology has better execution and throughput in correlation with other work in the field. The precision is improved from 81% to 93% in our work.

Singh et al(2020). designed software based on classification algorithms (including logistic regression, random forest, and naive Bayes) to predict the risk of liver disease from a data set with liver function test results

Vijayarani and Dhavanand (2016) found that SVM performed better over naive Bayes to predict cirrhosis, acute hepatitis, chronic hepatitis, and liver cancers from patient liver function test results.

The “Bioinformatics” is the area of research, which has been invented by two great researchers Ben Hesper and Paulien Hogeweg(2011). The invention of the term “Bioinformatics” had been created in the early stage of the 1970s. The two researchers Ben Hesper and Paulien Hogeweg used the term “Bioinformatics” in their research work. In this research work, the term bio informatics has been defined as “the study of informatics processes in biotic systems” .Further modification of the Bioinformatics had been done in the year 1981.

Kellis, (2014).Protein structure prediction is one of the finest sections of research work in the field of bioinformatics. The human genes are generated by the proteins. Proteins are the chain of amino acids. The protein structures are two types – primary protein structures and secondary protein structure. The primary protein structures are easy to be predicted because the primary protein structure is three dimensional. However, the secondary protein structures are higher than the three dimensional. Therefore, it is very complicated to determine protein structure of the secondary section. Mainly, these proteins are quaternary in structure. To predict the secondary protein structure, biometric tool can be used with the help of crystallography

Breda,( 2007).The genetic structure is the sequence of human genes. The genes are the building block of any living organisms. Sequence Analysis is the process of understanding the structure, functionalities, and features of the gene structure. Computer Science is very much helpful in the process of Sequence analysis. Computer Science offers several tools, which are very much powerful to analyse the sequence of genomes. Each of the such tools has some advantages and disadvantages. These tools have been applied on the merit of problems addressed. These tools are capable of identify the mutation of protein structure in the DNA of any organisms. Shotgun sequence technique is the best example of tools used for sequence analysis.

According to Gogi and Vijayalakshmi (2018), a prognosis of LD was detected using machine learning techniques. For detecting LD LFT dataset was used that has 11 attributes. In the research paper, 5 data mining classification techniques were used and the platform was MATLAB2016. The accuracy found for linear discriminant algorithm was 95.8 % and ROC was 0.93.

In the research paper, Midhila et al. (2017) described a computer-based analysis and classifications for detect 10 types of LD from ultrasound images using some techniques such as segmentation despeckling, feature extraction and gray level difference weights method. The accuracy of classification and segmentation in detecting cysts was 90% and 80% respectively.

Kumar and Katyal (2018) briefs a method for analyzing LD using data mining techniques. In this research paper, they created a classifications model for diagnosis and to forecast liver problems using 5 data mining algorithms and 1 boosting algorithm. Without boosting, the method's best accuracy of 72.18% was found.

Spann et al. (2020) explained a comprehensive review about LD and transplantation based on machine learning approaches. In the review paper, the authors found that if the patient's data are too large supervised machine learning tools can detect nonalcoholic fatty liver disease (NAFLD) at an early stage.

In their research model, Ramkumar et al. [9] depict liver cancer prediction based on conditional probability Bayes theorem. In the research WEKA tools and data mining techniques were used to predict liver cancer. It is found that drinking alcohol caused LD based on Bayes theorem.

Kefelegn and Kamat (2018) presented a survey that used data mining methods to predict and analyze liver problem diseases. Three algorithms such as Naïve Bayes, SVM and C4.5 have been utilized in the study approach. The model has evaluated the performance utilizing a confusion matrix and 10-fold cross-validation for the partitioning of data

Andrade et al. (2012) used extracted features of ultrasound images for diagnosing liver steatosis by comparing the performance of three different classifiers.

Automatic ROI selection and hierarchical method were used by Ojwimehr et al. (2017)for classifying normal persons from persons who are suffering from fatty liver diseases like steatosis, fibrosis, and cirrhosis.

Support vector machine was used by Li et al. (2008) for diagnosing and analyzing B-mode ultrasonic images of FLD with different textures.

The Bayesian framework was used by Ribeiro et al. (2009)for extracting features and for classifying normal liver ultrasonic images from fatty liver ultrasonic images where the anatomic and echogenic information of ultrasonic images were used.

The SVM classifier was used by Subramanya et al. (2015) in 2014 for classifying liver images which were from the US liver image dataset and they achieved an accuracy of 84.9%.

An accuracy of 81.2% was achieved by Ma et al.(2016) in 2015 by developing a signal processing approach that was based on kurtosis scanning and grading the fatty liver diseases and this was implemented on the US liver image dataset.

Backpropagation Neural Network (BPNN) was used by Saba et al.(2016)in 2016 where BPNN had 10 hidden layers and 128 features were used for extraction on US liver images by using different feature extraction algorithm total 6 different types of algorithms was used and BPNN gained accuracy of 97.6%.

Vijayrani et al. (2015) used the ILPD dataset which was collected from UCI and they implemented Support vector machine and Naïve Bayes classification algorithms for classification where they used MATLAB for implementation and according to the result SVM performed good gained high accuracy than Naïve Bayes which was 79.66% but Naïve Bayes took less time for training when compared to SVM

In 2014 Dhamodharan et al. (2014) used the Naïve Bayes and F1 tree algorithm for the prediction of liver diseases where three main liver diseases were classified as Liver Cancer, Cirrhosis, and Hepatitis where accuracy was used for selecting the best suitable algorithm where Naïve Bayes performed better than other algorithms.

Ramana(2011). made a critical study on liver diseases diagnosis by evaluating some selected classification algorithms such as naïve Bayes classifier, C4.5, back propagation neural network, K-NN and support vector. The authors obtained 51.59% accuracy on Naïve Bayes classifier, 55.94% on C4.5 algorithm, 66.66% on back propagation neural network, 62.6% on KNN and 62.6% accuracy on support vector machine. The poor performance in the training and testing of the liver disorder dataset as resulted from an insufficient in the dataset .

We, have also gone through a research paper Diagnosis of Liver Disease Using Machine Learning Techniques by Joel Jacob1 et al(2018).They have used FOUR classification algorithms Logistic Regression, Support Vector Machines (SVM), K Nearest Neighbor (KNN) and artificial neural networks (ANN) have been considered for comparing their performance based on the liver patient data. Authors obtained 73.23% accuracy on Logistic Regression, 72.05% on k-NN, 75.04 accuracy on Support vector machine.

We have also gone through a paper Liver Patient Classification using Intelligence Techniques by Jankisharan Pahareeya,et al(2014)In this paper Authors have used six intelligence techniques on

# 2.3. LITERATURE REVIEW

The healthcare landscape is undergoing significant transformation, driven by technological advancements and the increasing need for efficient and accurate disease prediction and management. Wang et al., [16] highlights the pressing need for fundamental reform in the American healthcare system since, despite a substantial budget, it lags behind peer-developed nations in terms of results like life expectancy. Using advanced machine learning methods such as Random Forest and Support Vector Regression (SVR) in conjunction with traditional statistical forecasting methodologies, the study projects future healthcare spending as a percentage of GDP for the year 2050. It's interesting to observe that the Random Forest and AutoRegressive Integrated Moving Average (ARIMA) models do similarly well in forecasting. The study underscores the critical role that healthcare analytics plays in comprehending the complexities of the healthcare system, in addition to underscoring the urgent need for appropriate policies to address the rising trajectory of healthcare spending and its effects on public health and the economy. The goal of the Yeganeh et al., [17] is to improve early detection of abnormalities in healthcare processes by introducing a Multistage Process Monitoring (MPM) tool designed specifically for healthcare data. The tool enhances detection capacities through the integration of machine learning approaches with statistical control charts. The MPM tool shows exceptional effectiveness in monitoring and guaranteeing patient safety through simulations and an actual case study on thyroid cancer surgery. Zini & Carcasci, [18] examines the energy use of an Italian hospital with particular attention on its Heating, Ventilation, and Air Conditioning(HVAC) system. Key energy drivers are identified through a methodical feature selection procedure, and artificial neural networks are used to estimate energy consumption. The study shows how the technique may identify unusual patterns in energy usage, offering a dependable and useful way to manage energy use in smart buildings.

# CHAPTER 3

# RESEARCH METHODOLOGY

# 3.1 Introduction

In this chapter, we present the methodology for liver disease prediction using machine learning techniques. Liver disease is a significant health issue worldwide, and early detection plays a crucial role in effective treatment and prevention. Machine learning algorithms offer the potential to analyze complex medical data and provide accurate predictions. In this chapter, we outline the system description, system architecture, system features, and the methodology employed for liver disease prediction. Additionally, we discuss data collection, data preparation, and system requirements. Furthermore, we recommend suitable algorithms for liver disease prediction based on their performance and interpretability.

# 3.2 System Architecture

The liver disease prediction system consists of several components that work together to achieve accurate predictions. The system architecture includes data preprocessing, feature selection, model training, and prediction generation. The input data is processed through these components, utilizing machine learning algorithms to build predictive models.

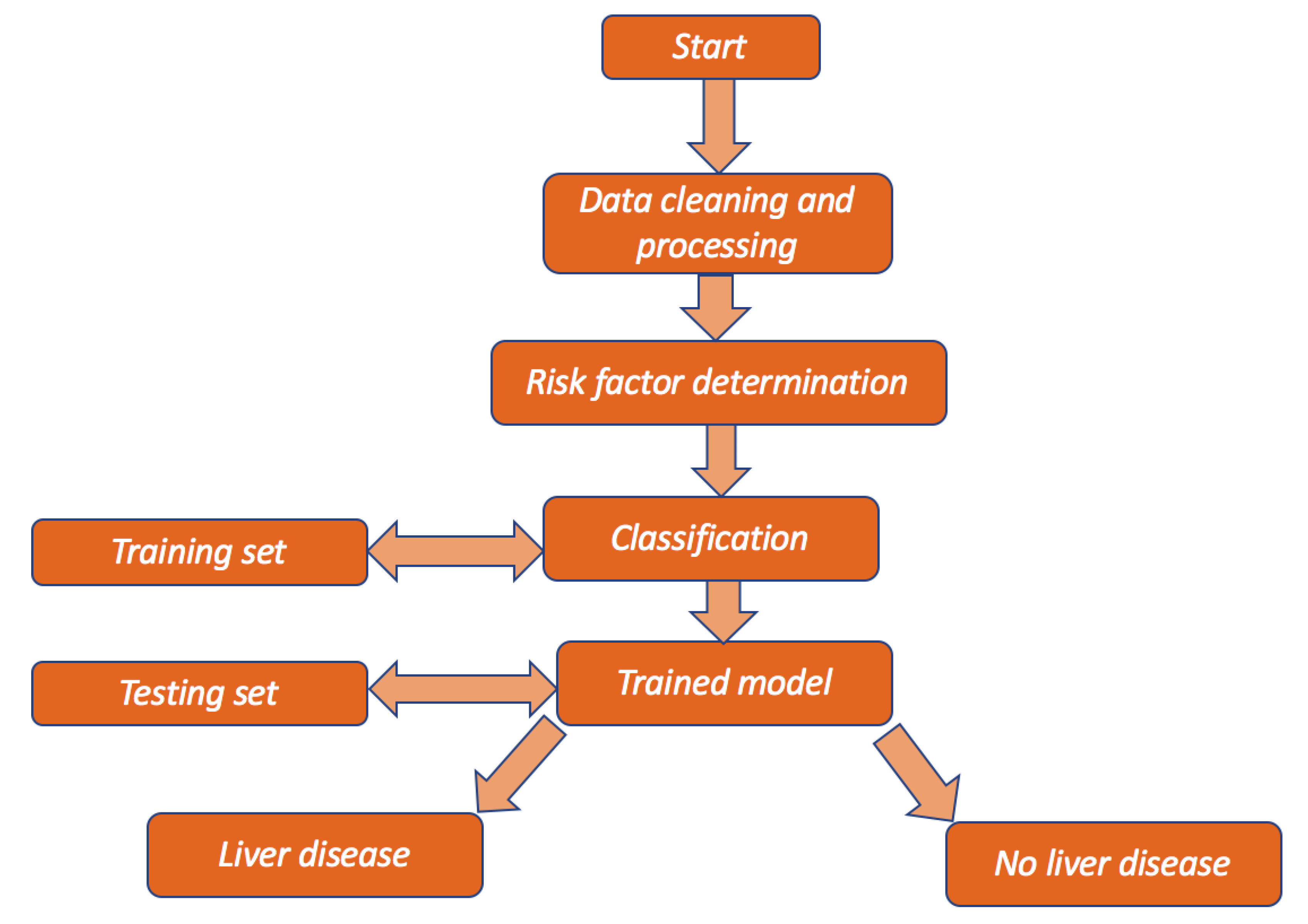
. 

Figure 3. 1 This diagram shows the system architecture of liver disease prediction using machine learning.

# 3.2.1 System Features

The liver disease prediction system offers several features to enhance the prediction process. These include:

User-friendly interface: The system provides an intuitive interface for healthcare professionals to input patient data and obtain predictions.

Data visualization: The system incorporates data visualization techniques to facilitate the understanding of patient data and model outcomes.

Model interpretability: The system employs machine learning algorithms that offer interpretability, allowing healthcare professionals to understand the factors influencing the predictions.

Real-time predictions: The system provides real-time predictions, enabling prompt decision-making and intervention.

Report: the system also print the reports of patient that enter data in textbox of input fields.

# 3.2.2 Methodology

The liver disease prediction methodology involves the following steps:

# 3.2.3Data Collection

For the liver disease prediction system, data collection is a crucial step. The dataset should

include records of patients diagnosed with liver diseases, along with relevant features such as age, gender, total bilirubin, direct bilirubin, alkaline phosphatase, alamine aminotransferase, aspartate aminotransferase, total proteins, albumin, and albumin/globulin ratio, and various laboratory test results. the success of any system based on machine learning is highly dependent on the quality and quantity of data used

for training. In this study, we get a dataset from Kaggle, and it contains 11 columns or

variables that make it easy for us to make predictions about liver disease. Data collection is the process of gathering and acquiring relevant data to be used for analysis, decision- making, or research purposes. It involves identifying the data needed, designing a data collection plan, collecting the data, and ensuring its quality and integrity.

The variables (also known as features or attributes) that are available for liver disease prediction are as follows:

Age: The age of the individual.

Gender: The gender of the individual (e.g., male or female). Total Bilirubin: The total amount of bilirubin in the blood.

Direct Bilirubin: The amount of bilirubin specifically conjugated with glucuronic acid.

Alkphos-Alkaline Phosphotase: The level of alkaline phosphatase, an enzyme found in various tissues including the liver.

Sgpt Alamine Aminotransferase: The level of alanine aminotransferase (ALT), an enzyme predominantly found in the liver.

Sgot Aspartate Aminotransferase: The level of aspartate aminotransferase (AST), an enzyme found in various tissues including the liver.

Total Proteins: The total amount of proteins in the blood.

ALB Albumin: The level of albumin, a protein synthesized by the liver.

A/G Ratio Albumin and Globulin Ratio: The ratio of albumin to globulin, which provides information about liver function and overall health.

# 3.2.4 Data Preparation

Data preprocessing is essential to ensure the quality and suitability of the data for machine learning algorithms.

**The main goals of our dataset preparation are:**

* **Handling missing data:** Identifying and addressing missing values in the dataset.

This can involve removing rows with missing values, imputing missing values based on statistical measures, or using advanced missing values based on techniques to predict missing values.

* **Encoding categorical variables:** converting categorical variables into numerical format that machine learning models can understand.
* **Data splitting:** dividing dataset into training and test sets. The training set is used to train the model. The testing set is used to evaluate to model’s performance of unseen data.
* **Handling data duplicates:** identifying and handling duplicate rows in dataset to avoid biases in model training.
* **Outlier:** An outlier (or outlier) is a data point that significantly differs from the other

observations in a dataset. Outliers can arise due to variability in the data, measurement errors, or experimental errors. They can be unusually high or low values that do not fit the general pattern of the data. Identifying and analyzing outliers is important because they can impact statistical analyses and machine learning models, potentially loading to skewed result and misleading conclusion.

# 3.2.5 Model Selection

In this step, suitable machine learning algorithms are selected for liver disease prediction. The choice of algorithms depends on factors such as the nature of the data, the size of the dataset, and the interpretability requirements. Commonly used algorithms for liver disease prediction include decision trees, random forests, logistic regression, support vector machines, and neural networks.

In this research we used five algorithms for using liver disease prediction such as logistic regression, random forest, decision trees, neural network and gradient boosting.

The following is how we use the model or algorithms and what each one does is mentioned in our study to predict liver disease:

**Logistic Regression**: Logistic regression is a statistical algorithm used for binary classification problems, where the dependent variable is categorical and has two classes (in this case, liver disease or no liver disease). It models the relationship between the independent variables (age, gender, total bilirubin, direct bilirubin, alkaline phosphatase, alamine aminotransferase, aspartate aminotransferase, total proteins, albumin, and albumin/globulin ratio) and the probability of belonging to a particular class.

Logistic regression uses a logistic function (also known as a sigmoid function) to map the linear combination of the independent variables to the range [0, 1], representing the probability of the positive class. The model learns the coefficients (weights) for each independent variable during the training process. These coefficients indicate the strength and direction of the relationship between the independent variables and the likelihood of having liver disease.

Logistic Regression is a suitable choice for binary classification problems like liver disease prediction. It can handle both numerical and categorical input variables. You can use age, gender, Total Bilirubin, Direct Bilirubin, Alkphos Alkaline Phosphotase, Sgpt Alamine Aminotransferase, Sgot Aspartate Aminotransferase, Total Protiens, ALB Albumin, and A/G Ratio Albumin and Globulin Ratio as features to train a logistic regression model.

To apply logistic regression, you would first prepare your dataset by splitting it into two parts: the feature matrix (X) and the target variable (y). The feature matrix (X) would contain the predictor variables (age, gender, etc.), and the target variable (y) would contain the binary labels indicating whether a person has liver disease or not.

Next, you would split your dataset into a training set and a test set. The training set is used to train the logistic regression model, while the test set is used to evaluate its performance.

Once you have the data split, you can train the logistic regression model using the training set. The model will learn the relationship between the input variables and the probability of having liver disease. It will estimate the coefficients (weights) for each input variable, indicating their contribution to the prediction.

During training, logistic regression applies a logistic function (sigmoid function) to the linear combination of the input variables and their corresponding coefficients. This transformation maps the linear output to a probability value between 0 and 1. The logistic function allows us to interpret the output as the probability of having liver disease.

After training, you can use the trained logistic regression model to make predictions on new, unseen data. Given the input variables of a new individual, the model will calculate the probability of that person having liver disease based on the learned coefficients.

To evaluate the performance of the logistic regression model, you can use metrics such as accuracy, precision, recall, and F1-score. These metrics will help you assess how well the model predicts the presence or absence of liver disease.

**Random Forest:** Random Forest is an ensemble learning algorithm that combines multiple decision trees to make predictions. It is widely used for both classification and regression tasks, including liver disease prediction.

**Here's how Random Forest works with your dataset:**

Dataset Preparation: First, you would split your dataset into a feature matrix (X) and a target variable (y). The feature matrix (X) would contain the predictor variables (age, gender, Total Bilirubin, etc.), and the target variable (y) would contain the binary labels indicating whether a person has liver disease or not.

Building the Random Forest: Random Forest consists of an ensemble of decision trees. Each decision tree is trained on a random subset of the training data, and the final prediction is made by aggregating the predictions of all the individual trees.

To build a Random Forest model, you would specify the number of decision trees (n\_estimators) and other hyperparameters such as the maximum depth of each tree and the number of features to consider at each split.

**Decision trees:** A decision tree is a popular machine learning algorithm used for both classification and regression tasks. It's a predictive model that maps observations about an item to conclusions about its target value. In your case, for predicting liver disease, a decision tree can be constructed to analyze various factors or features that might influence the presence or absence of liver disease in an individual.

Here's a basic explanation of how decision trees work in the context of liver disease prediction:

Tree Structure: A decision tree is a hierarchical structure consisting of nodes and branches. Each internal node represents a "decision" based on a feature, and each branch represents the outcome of that decision. Leaf nodes represent the final outcome, which in this case would be the prediction of whether the individual has liver disease or not.

Splitting Criteria: At each internal node, the decision tree algorithm selects the best feature to split the data. The goal is to create homogeneous subsets of data regarding the target variable (presence

Prediction: Once the tree is constructed, it can be used to make predictions for new, unseen data. To classify a new instance , the algorithm traverses the tree from the root node down to a leaf node based on the values of the features. The prediction at the leaf node is then assigned to the instance.

In the context of liver disease prediction, the decision tree would be trained on a dataset containing various features such as age, gender, Total Bilirubin, Direct Bilirubin, Alkphos Alkaline Phosphotase, Sgpt Alamine Aminotransferase, Sgot Aspartate Aminotransferase, Total Protiens, ALB Albumin, and A/G Ratio Albumin and Globulin Ratio.

**Gradient Boosting Algorithms:** Gradient boosting algorithms, such as XGBoost or LightGBM, are ensemble learning methods that build an ensemble of weak models (typically decision trees) in a sequential manner. They aim to improve the performance by iteratively correcting the mistakes of the previous models.

Gradient Boosting is a powerful machine learning algorithm that can be used for liver disease prediction with the dataset we have include age, gender, Total Bilirubin, Direct Bilirubin, Alkphos Alkaline Phosphatase, Sgpt Alamine Aminotransferase, Sgot Aspartate Aminotransferase, Total Proteins, ALB Albumin, and A/G Ratio Albumin and Globulin Ratio.

**Artificial Neural Networks (ANN):** Artificial Neural Networks (ANNs) are machine learning models inspired by the structure and function of the human brain. They consist of interconnected nodes (neurons) organized in layers. ANNs can have multiple hidden layers, allowing them to learn complex patterns and relationships in the data.

In the case of liver disease prediction, ANN can be trained on the dataset we sing, with each input neuron representing a specific feature (age, gender, total bilirubin, direct bilirubin, alkaline phosphatase, alamine aminotransferase, aspartate aminotransferase, total proteins, albumin, and albumin/globulin ratio). The hidden layers perform computations on the input features, and the output layer produces the prediction. During training, the weights connecting the neurons are adjusted to minimize the difference between the predicted and actual outputs.

ANNs are known for their ability to capture complex patterns and relationships in the data. However, they require a large amount of training data and computational resources for training. Interpretability can also be a challenge with ANNs, as they are often considered black-box models.

# 3.3 Model Training

The selected algorithms are trained using the prepared dataset. The dataset is divided into training and testing sets, and the models are trained on the training set. The performance of the models is evaluated using appropriate evaluation metrics such as accuracy.

Start by splitting your dataset into a feature matrix (X) and a target variable (y). The feature matrix should contain the predictor variables, such as age, gender, Total Bilirubin, etc., while the target

To evaluate the performance of your model, it's common practice to split your dataset into training and test sets. The training set is used to train the model, while the test set is used to evaluate its performance. This ensures that the model is tested on unseen data to assess its ability to generalize.

# 3.3.1 Prediction Generation

Once the models are trained and evaluated, they can be used to generate predictions on new, unseen data. The input data is processed through the trained models, and the predictions are generated based on the learned patterns and relationships. The predictions can be presented to healthcare professionals through a user-friendly interface, providing valuable insights for diagnosis and treatment decision-making.

# 3.3.2 System Requirements

The hardware and software requirements for building the project of liver disease prediction system using machine learning vary depending in your operation system and digital platform

For which you are developing this system there are some requirements. These are hardware and software requirements, respectively, and they should be available when developing

this project. Although these requirements depend on how complex the system is even the least complex one also has some requirements for you to develop that particular software

# 3.3.3 Hardware Requirements

A computer system is made up of units that are put together to work as one to achieve a common goal. The hardware device requirements for the implementation of Liver disease prediction system using machine learning are:

# Table 3.1 Hardware Requirements

Table 3. 1 Hardware Requirements

|  |  |
| --- | --- |
| DEVICE | Description |
| Processorc | Intel(R) Core (TM) i5-4300U CPU @ 1.90GHz (4 CPUs), 2.5GHz |
| RAM | 8GB Or Above 8GB |
| Display | 2.7inch |
| System Type | 64-bit Operating System |
| Operating System | Windows 10 or Window 11 Pro |

# 3.6.2 Software Requirements

This section explains the minimum software requirements for using our program, as shown in the table below.

Table 2: Software Requirement’s

|  |  |
| --- | --- |
| Software | Minimum requirement |
| Operating System | Windows 10 or Window 11 Pro |
| Python flask | Python 3.11.9 |
| Microsoft Excel | Excel Version 2019 |
| Jupyter-notebook | 7 |
| Mongo-db | 7.0 |

# CHAPTER 4

# RESULTS AND DISCUSSION

## 4.1 Introduction

This chapter focuses on result and discussion the liver disease prediction using machine learning techniques. It presents the project's practical aspects, including the system's development and the evaluation of its performance through comprehensive testing.

In this also chapter, the implementation details of the machine learning models developed for liver disease prediction are presented. This chapter also encompasses the testing phase, where the efficacy of the models is assessed using appropriate evaluation metrics.

## 4.2 Discussion

The problem we solved was to create a liver disease prediction system using machine learning.

The main goal of our research was to create or to develop accurate and reliable machine learning models for predicting liver diseases. By leveraging advanced algorithms and comprehensive datasets, researchers aim was to create models that can effectively identify risk factors, predict disease progression, and inform personalized treatment plans. Such models hold the promise of revolutionizing liver disease management, leading to improved patient outcomes, reduced healthcare costs, and advancements in medical research. Furthermore We want to create an easy and visually appealing user interface that allows users to easily input their medical history of liver disease based the dataset we sign.

We learned from this study that liver disease can be recommended using machine

learning. We learned a lot about how the machine works and the algorithms of the machine. One of the biggest challenges we faced was the collection of data, which we faced a lot of difficulties, because there was no data on liver disease in Somalis, if we found it, it was very difficult for us to collect that data from the hospitals in our country. Because the hospitals in our country do not have

## 4.3 Overview of the implementation environment

The main objective of this study is to build a system that predict liver disease in patients to know whether the patient is healthy in his liver and not healthy in his liver.

Our system for implementing machine learning and webserver. The graphical user component is the Python flask at the front end. We also use Python so that the Models of Data Collection program must be implemented Dataset as the back-end.

We also use a database connection called mongodb to store the data entered users who use our website to receive reports.

## 4.4 Data Preprocessing

Before model implementation, the raw data underwent preprocessing steps to ensure its suitability for machine learning analysis. These steps included data cleaning, handling missing values. Detailed explanations of each preprocessing technique are provided, along with justifications for their application in the context of liver disease prediction.

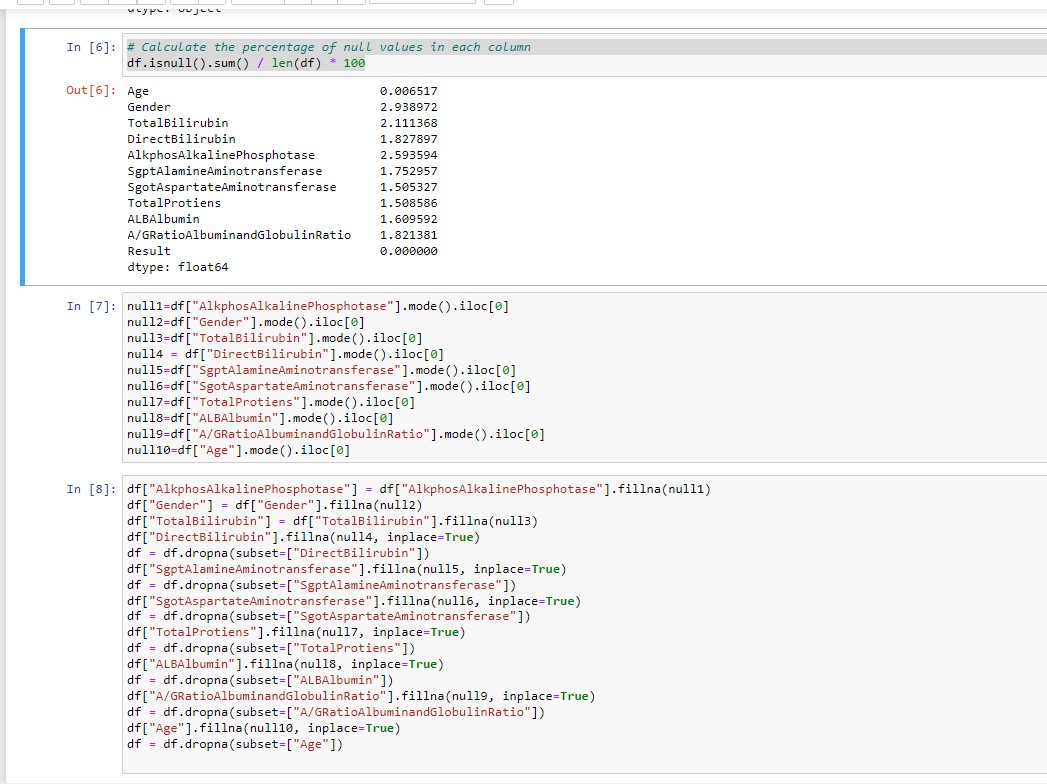


Figure 4. 2 **mission values of null value**



Figure 4. 3 **mission values of duplicate**

#### 4.1Training and Testing

After training and evaluating five different models on our dataset, we found notable variations in their performance metrics. Among these models, gradient boost model for training we got accuracy of 97.77 and gradient boost model for testing we got accuracy of 97.65.

These results underscore the importance of considering the nature of the data when selecting a suitable model, as well as the potential impact of feature types on predictive performance

This accuracy value indicates that the Gradient Boosting model achieved an accuracy of 97.77% on the training dataset. high training accuracy suggests that the model was able to effectively learn and capture the patterns present in the training data. However, it's important to note that high training accuracy does not guarantee good performance on unseen data, as the model might have memorized the training examples rather than learning generalizable patterns.

This accuracy value indicates that the same Gradient Boosting model achieved an accuracy of 97.65% on the testing dataset, which represents unseen data that the model was not trained on.

The gap between training and testing accuracy is relatively small, indicating that the model's performance is consistent across both datasets.

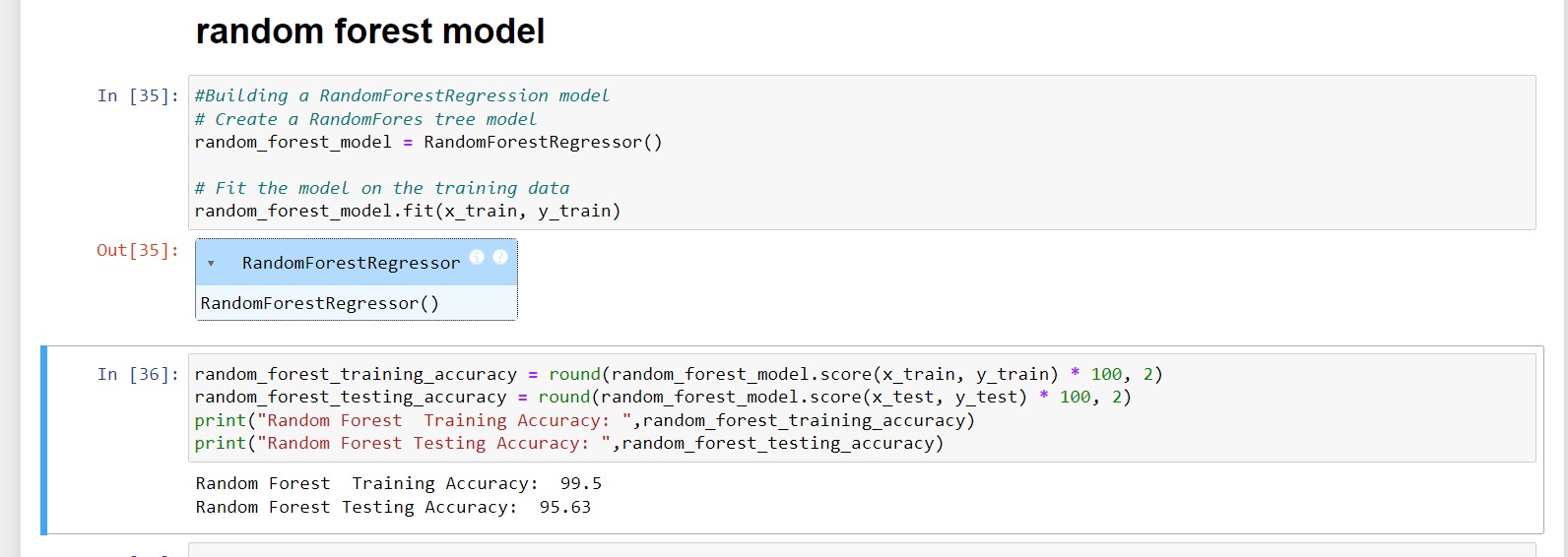


Figure 4. 4: Random Forest for Classification Training testing accuracy

The testing accuracy being slightly lower than the training accuracy is expected in most cases and is a good sign that the model is not overfitting excessively. Achieving high accuracy on the testing dataset suggests that the model generalized well to new, unseen data and made accurate predictions.

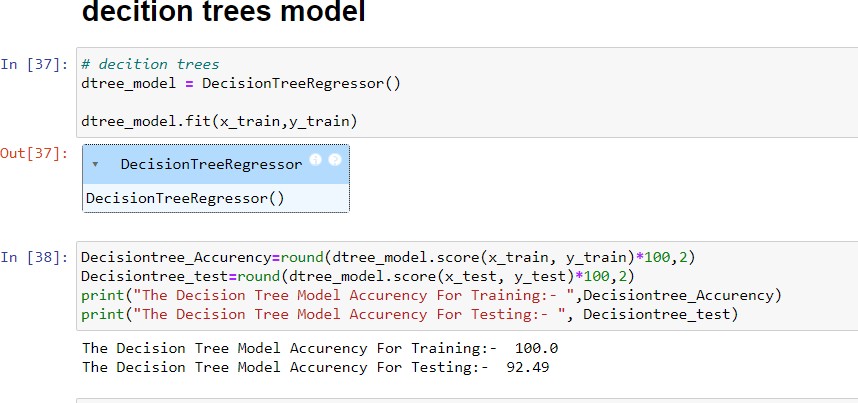


Figure 4. 5: Decision trees model for Classification Training testing accuracy

## MODEL SELECTION

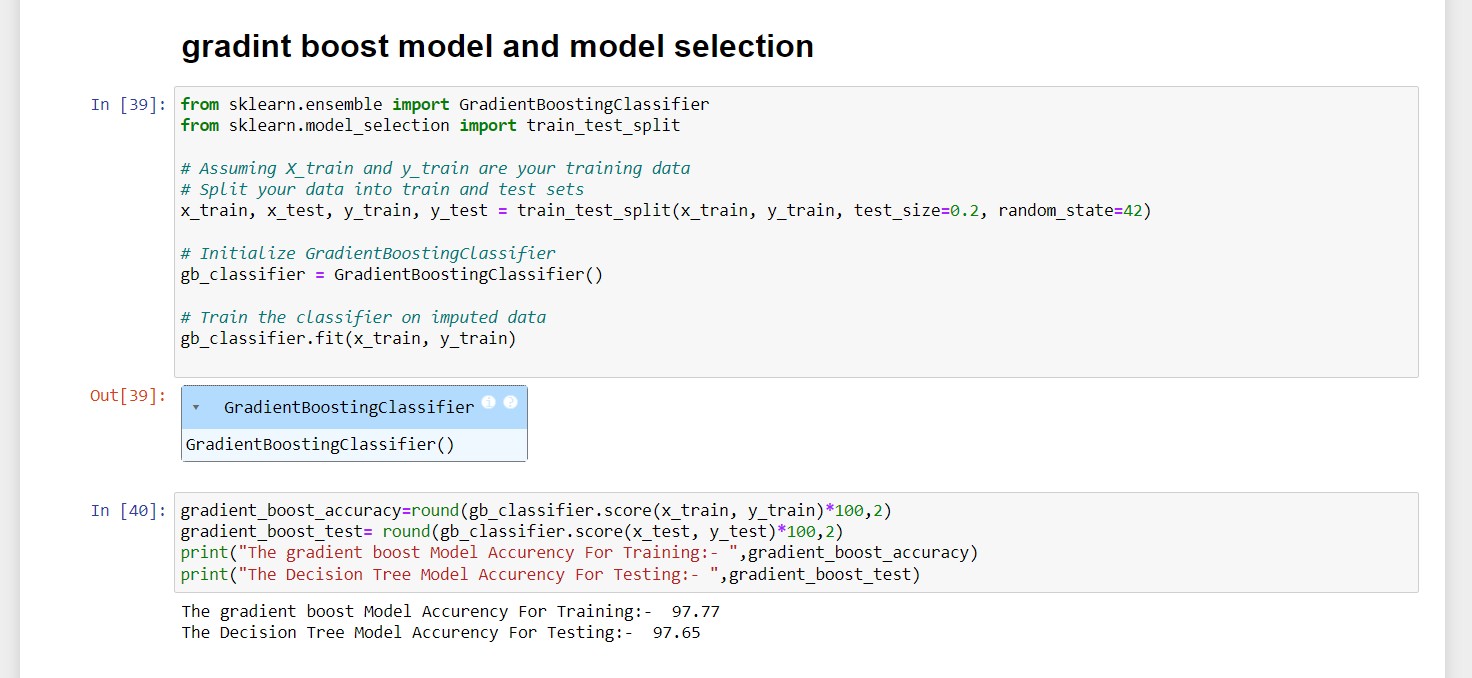
we selected this model gardient boost model

Figure 4. 6 Model selection

**Breid description:**

We selected the gradient-boost model to predict the liver disease prediction using machine learning. The gradient-boost model for liver disease prediction can enhance user engagement, and satisfaction and make suitable prediction of liver disease prediction using machine learning.

The performance of this model is good because the training accuracy and testing accuracy of that model is very good. training accuracy and test accuracy are very close and we found very similar results when we analyzed the data

## 4.5 Snapshots of the system

This system has two major components one is Fort-end and back-end of each section we will explain it separately and include her pictures

## 4.5.1 Front-end

The area of web development that focuses on what users view on their end is known as front-end development. It requires transferring the code created by back-end developers into a graphical interface and ensuring that the data is displayed understandably. All you'd see on a website or web application without Front End Development are unreadable scripts. People without coding experience, on the other hand, may quickly understand and use web applications and webpages due to Front-End developers. Everything you see on Google Apps, Canva, Facebook, and other web services is the result of collaboration between back- end and front-End Engineer.

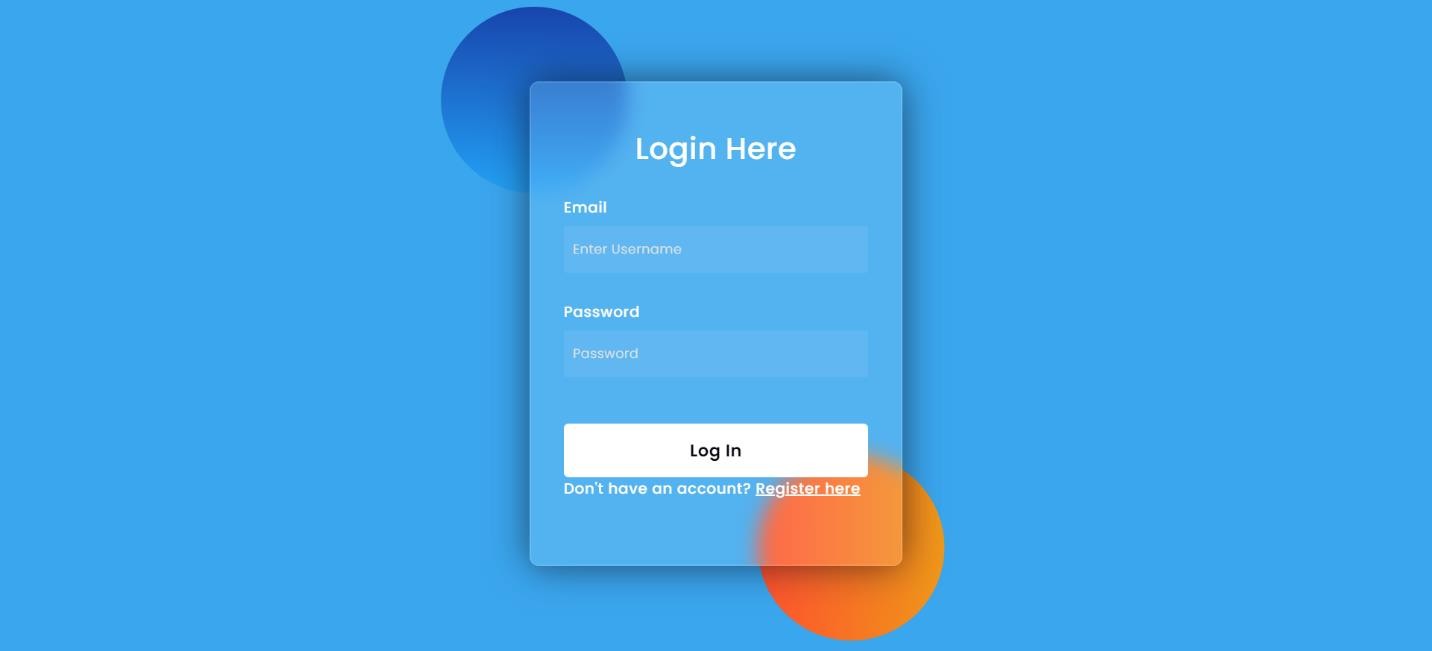


Figure 4. 7: Login page

**Breid description**:

A login page is a web page that allows users to authenticate themselves in order to gain access to a secure area or specific features of a website or web application. It typically consists of fields where users can input their credentials, such as a username or email address and a password. Once the user submits this information, it is typically validated against a database of registered users to confirm their identity. If the credentials are correct, the user is granted access to the restricted areas or features of the website.

Login pages are essential for websites that require user authentication to provide personalized experiences, access to private information, or to perform certain actions. They help ensure the security and privacy of user accounts by verifying their identity before granting access to sensitive data or functionalities.

The login page has an email and password that the person who wants to use the website should enter the email and password. We read the user from the database, if the email or password is wrong or does not exist, we will catch the error.

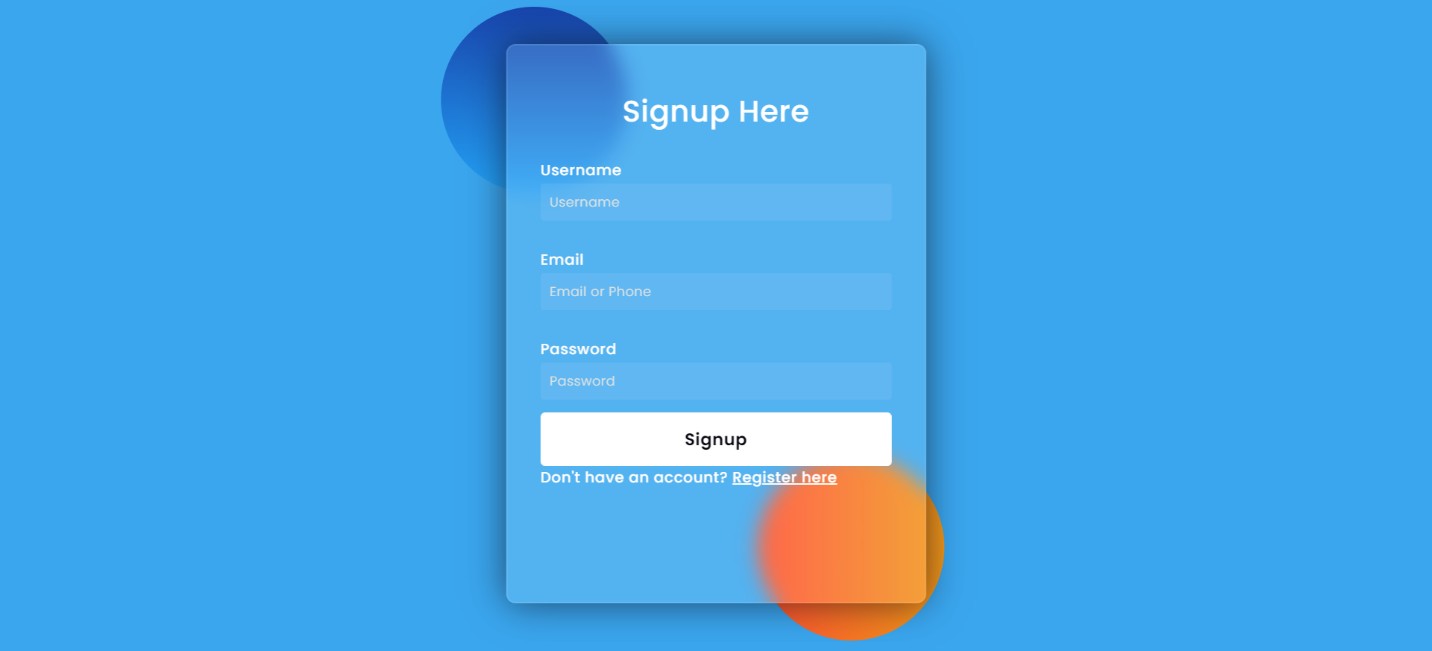


Figure 4. 8: Signup Page

**Breif description:**

A signup page, also known as a registration page, is a web page where users can create a new account to access a website or web application. It's essentially the entry point for users who are new to the platform and want to become members. Here's a breakdown of the typical components you might find on a signup page:

Registration Form: This is the central element of the signup page where users input their information to create an account. It usually includes fields for essential information such as:

Username: The unique name that identifies the user within the system.

Email Address: A valid email address that will be associated with the user's account.

Password: A secure password to protect the user's account. It's usually required to meet certain criteria for strength.

Confirm Password: A repeat of the password to ensure it was entered correctly.

"Sign Up" Button: Clicking this button submits the information entered in the registration form for account creation.

Error Messages: If users input invalid information or encounter any issues during the signup process, error messages should appear to guide them in correcting their mistakes. These messages might include notifications such as "Username already taken" or "Invalid email address."

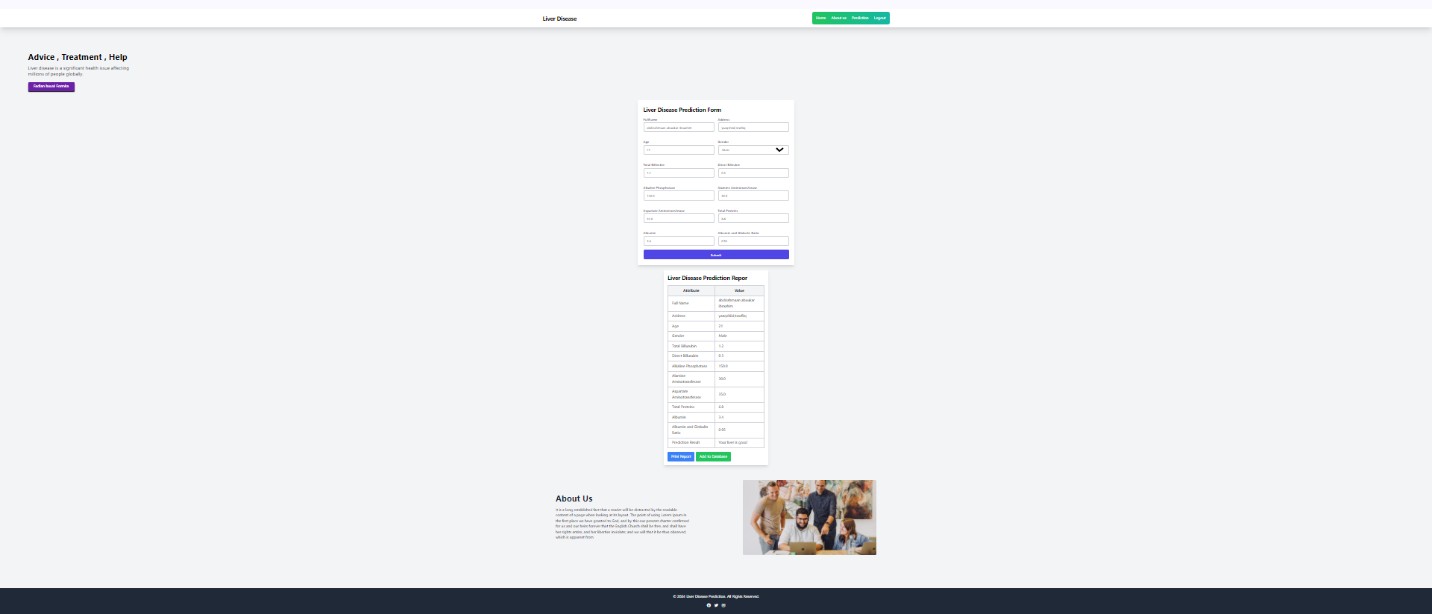


Figure 4. 9: Home Page

**Brief description:**

A home page is the main or initial webpage of a website. It serves as the starting point for visitors to navigate through the site's content. Typically, the home page contains an overview of what the website offers, navigation links to other pages within the site, and possibly some featured or important content. It's often designed to provide easy access to the most relevant information or services offered by the website. Home pages are usually the first page users see when they visit a website, and they often set the tone and provide an impression of the site's purpose and content.

Also on the home page there is a prediction form which is the main part of the system and it contains input fields where the user enters data and after the user enters the data the machine makes a prediction based on the dataset we have trained with machine learning or the model we have chosen. which is a gradient-boost model. We also save the data that the user enters in the database that we have connected to the system when the user clicks on the submit or predict button. the database we used is mongodb.

## 4.5.2 Back-end

The back end refers to the sections of the code that allow it to perform but are not visible to the user. The back end of a computer system stores and accesses the majority of data and operational procedures. One or more programming languages are usually used in the code. The back end, often known as the data access layer of software or hardware, contains any functionality that requires digital access and control.

Flask is primarily used as a backend framework for building web applications. It's a lightweight and flexible Python web framework that provides tools, libraries, and functionalities to develop server-side logic, handle HTTP requests, and interact with databases. While Flask itself is mainly focused on the backend, it can be used in conjunction with frontend technologies like HTML, CSS, and JavaScript to create full- stack web applications. Flask serves as the backend that handles the business logic, data processing, and communication with the client-side, while the frontend technologies handle the presentation and user interaction aspects of the application. Also we used database connection called mongo-db to store the data that enter our users. We connect in flask framework..



Figure 4. 10: import necessary

**Brief description:**

This part is the imports section and it is done to import what it needs system. which is used to import necessary libraries and modules into the Python script.

In this Python script, we'll explore various machine learning models for both classification and regression tasks using the scikit-learn library along with some additional tools for data preprocessing and evaluation. We'll utilize techniques such as logistic regression, random forest, decision trees, k-nearest neighbors, gradient-boost and SMOTEENN for handling imbalanced data.

#### 4.5.3 Data Visualization

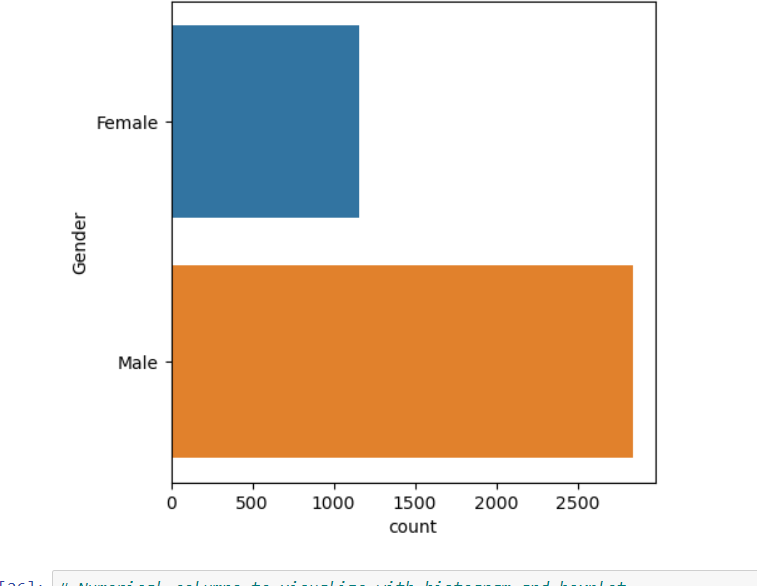


Figure 4. 11: Gender Visulization

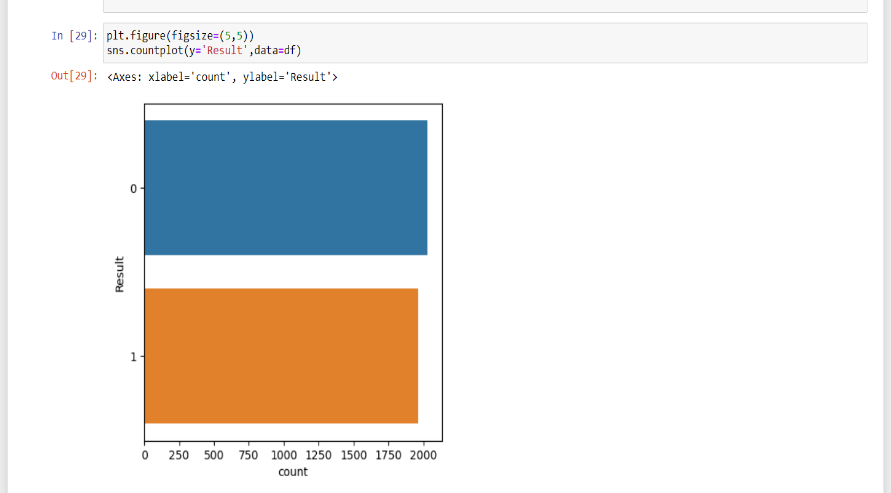


Figure 4. 12: Result Visulization

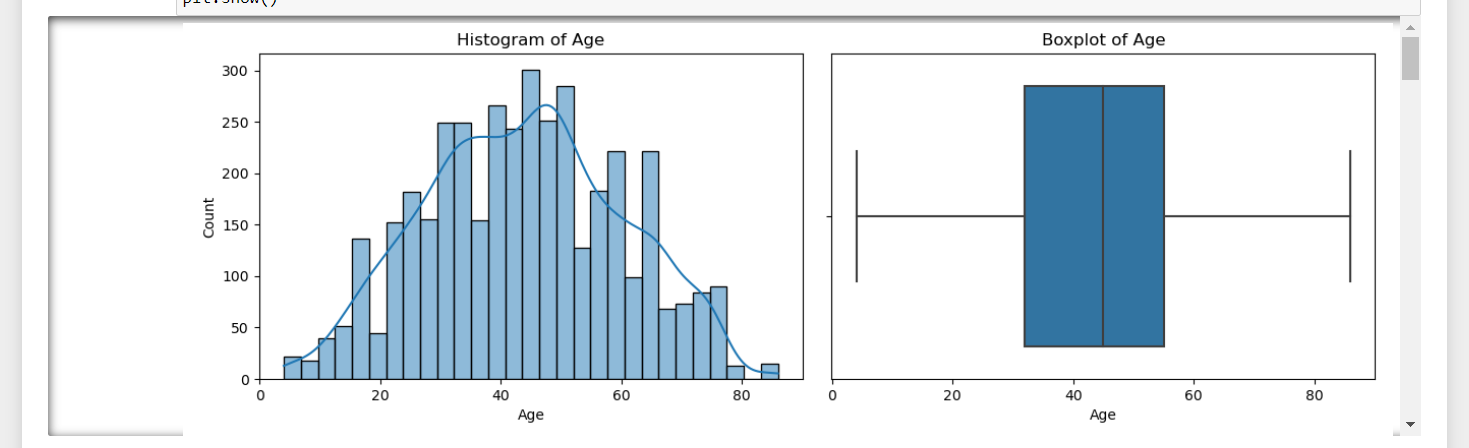


Figure 4. 13: Age Distribution

**Brief description:**

The image contains a Histogram and a Boxplot, both depicting the age distribution in a dataset. The histogram shows a bell-shaped age frequency, suggesting a roughly normal distribution, while the boxplot indicates the median age and the spread of the data without any visible outliers. Both graphs use the same age scale for a consistent comparison of the data.

# CHAPTER 5

# CONCLUSION AND LESSONS LEARNED

## 5.1 Introduction

In this chapter, we will talk about the conclusion of our research that we have been doing for a long time. Our research is about creating a system to predict liver disease by using machine learning. I will also talk about the future work of researchers who will complete the work There is our system and the things we have achieved and we have not had time.

## 5.2 Conclusion

The purpose of our research was to create or develop a system that makes us predict liver disease using machine learning techneque. when we did or solved this problem we faced many challenges and one of them was getting the dataset which was really difficult for us to get information from our hospitals in our country Somalia, we didn't have enough reliable data collection so we got the information from the kaggle.com. after that we trained our dataset and removed any missing value like null, duplicate and outlaier then we created the system and we used 5 models and we chose one model which is gradient boost as we mentioned chapter 6 and chapter 5 this model was the best for us After that, we got different results and we developed our system which is the best for us. We also created a website which is well designed and allows our users to easily use our system to know whether their liver is good or not. our study represents a significant step forward in leveraging machine learning for liver disease prediction. By building on these findings and addressing the outlined recommendations and future research directions, we can further advance the field and ultimately improve patient outcomes in the management of liver disease.

Collaboration with healthcare providers and consideration of societal impact will be crucial in ensuring the broader benefits of these systems in improving overall health outcomes. Ultimately, The main goal of our research was to create or to develop accurate and reliable machine learning models for predicting liver diseases.

## 5.3 Lessons Learned

Developing the liver disease prediction system using machine learning provided valuable insights, both technical and practical. One of the most important lessons learned was the value of using real-world health data. Due to the limited availability of public datasets from Somali healthcare institutions, we relied on open datasets such as the Indian Liver Patient Dataset (ILPD). While this allowed us to build and evaluate models effectively, it also highlighted the limitations of using non-local data. The absence of Somali-specific clinical trends and risk factors meant the model might not fully capture local healthcare dynamics. This emphasized the critical need for locally sourced and contextual medical data for future accuracy and trustworthiness in predictions.

Another key learning was the importance of data preprocessing in healthcare machine learning applications. The original dataset contained missing values, inconsistent entries, and imbalanced class distributions between liver disease and non-disease cases. Addressing these issues through imputation, normalization, and resampling techniques (like SMOTE) greatly improved model reliability. This process reinforced the idea that clean, balanced data is just as vital as algorithm selection.

On the modeling side, experimenting with different algorithms—such as Logistic Regression, Random Forest, and XGBoost—taught us that no single model fits all medical prediction problems. While tree-based models like Random Forest offered high accuracy and robustness to noisy data, simpler models like Logistic Regression provided better interpretability—critical for clinical decision-making. The project emphasized the importance of balancing model performance with explainability, especially in sensitive domains like healthcare.

We also learned the necessity of user-centered design when building interfaces for healthcare workers or patients. In low-resource settings, access to powerful hardware or stable internet may be limited. As a result, the system interface was designed to be lightweight, mobile-friendly, and intuitive, ensuring that medical professionals could use it easily during consultations. This design approach reminded us that technical functionality must always serve real-world usability, especially in health tech.

Another major insight was the need for secure and ethical handling of medical data. Even though this version used open-source data, the process of simulating deployment environments raised awareness about user privacy, data encryption, and regulatory compliance. Any future versions that use real patient data will require strict adherence to data protection laws and healthcare standards like HIPAA or GDPR.

From a backend perspective, integrating the trained model (.pkl) into a Django-based web application taught us how deployment differs from offline experimentation. We encountered challenges related to model loading time, server memory management, and inference latency. Solving these issues involved optimizing file storage paths, pre-loading the model, and caching predictions to ensure a smooth user experience in a real-time web environment.

## 5.4 Future Work

While our study has made significant strides in liver disease prediction using machine learning, several avenues for future research warrant exploration:

**Exploration of Novel Biomarkers**: Future research should investigate the utility of novel biomarkers and imaging modalities for enhancing the predictive accuracy of liver disease models. Emerging technologies such as metabolomics and imaging biomarkers hold promise for improving early detection and risk stratification.

**Personalized Medicine Approaches**: There is a growing need for personalized medicine approaches in liver disease management. Future research should focus on developing personalized predictive models that account for individual variability in disease progression, treatment response, and genetic predisposition.

**Longitudinal Studies and Real-Time Monitoring**: Longitudinal studies tracking patients over extended periods can provide valuable insights into the progression of liver disease and the effectiveness of interventions. Real-time monitoring technologies and wearable devices offer opportunities for continuous data collection and remote patient management.

**Ethical and Regulatory Considerations**: As machine learning applications in healthcare continue to evolve, it is essential to address ethical and regulatory considerations surrounding data privacy, consent, and algorithmic bias. Future research should prioritize ethical guidelines and regulatory frameworks to ensure the responsible and equitable deployment of predictive models in clinical practice.

## REFERENCE

Abdalrada, A. S., Yahya, O. H., Alaidi, A. H. M., Hussein, N. A., Alrikabi, H. T., & Al-Quraishi, T. A.-Q. (2019). A predictive model for liver disease progression based on logistic regression algorithm. Periodicals of Engineering and Natural Sciences, 7(3), 1255–1264.

Abdelghany, A., & Guzhva, V. (2010). A time-series modelling approach for airport short-term demand forecasting. Journal of Airport Management, 5, 72–87.

Abdella, J. A., Zaki, N., Shuaib, K., & Khan, F. (2019). Airline ticket price and demand prediction: A survey. Journal of King Saud University - Computer and Information Sciences.

Abdullah, S., & Rajalaxmi, R. R. (2012). A data mining model for predicting the coronary heart disease using random forest classifier. In Proceedings of the International Conference on Recent Trends in Computational Methods, Communication and Controls (pp. 22–25).

Afrin, S., Shamrat, F. J. M., Nibir, T. I., Muntasim, M. F., Moharram, M. S., Imran, M., & Abdulla, M. (2021). Supervised machine learning based liver disease prediction approach with LASSO feature selection. Bulletin of Electrical Engineering and Informatics, 10(6), 3369–3376.

Akbugday, E., et al. (2019). Performed classification on Breast Cancer Dataset by using KNN, SVM and achieved accuracy of 96.85%.

Aleksandrovich, P. V., Leopoldovich, K. I., & Viktorovich, P. A. (2018). Predicting sales prices of the houses using regression methods of machine learning. In 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC). DOI: 10.1109/RPC.2018.8482191.

Alkeshuosh, A. H., Moghadam, M. Z., Al Mansoori, I., & Abdar, M. (2017). Using PSO algorithm for producing best rules in diagnosis of heart disease. In Proceedings of the International Conference on Computer and Applications (ICCA) (pp. 306–311).

Al-milli, N. (2013). Backpropagation neural network for prediction of heart disease. Journal of Theoretical and Applied Information Technology, 56(1), 131–135.

Altaf, I., Butt, M. A., & Zaman, M. (2022). Disease detection and prediction using the liver function test data: A review of machine learning algorithms. [Conference proceedings, pp. 785–800; specific conference not specified.]

Andrade, A., Silva, J. S., Santos, J., & Belo-Soares, P. (2012). Classifier approaches for liver steatosis using ultrasound images. Procedia Technology, 5, 763–770.

Aneeshkumar, A. S., & Venkateswaran, C. J. (2012). Estimating the surveillance of liver disorder using classification algorithms. International Journal of Computer Applications, 57(6), 0975–8887.

Anooj, P. K. (2012). Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. Journal of King Saud University - Computer and Information Sciences, 24(1), 27–40. DOI: 10.1016/j.jksuci.2011.09.002.

Arbain, A. N., & Balakrishnan, B. Y. P. (2019). A comparison of data mining algorithms for liver disease prediction on imbalanced data. International Journal of Data Science and Advanced Analytics, 1(1), 1–11.

Choudhary, R., Gopalakrishnan, T., Ruby, D., Gayathri, A., Murthy, V. S., & Shekhar, R. (2021). An efficient model for predicting liver disease using machine learning. In Data Analytics in Bioinformatics: A Machine Learning Perspective (pp. 443–457).

D’Arcy, B., Gallagher, C., & Madden, M. G. (2015). A Bayesian classification approach to improving performance for a real-world sales forecasting application. In 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA). DOI: 10.1109/ICMLA.2015.150.

Dhamodharan, S. (2014). Liver disease prediction using Bayesian classification. COMPUSOFT: An International Journal of Advanced Computer Technology, Special Issue, 4th National Conference on Advanced Computing, Applications & Technologies, 1–3.

Ding, H., Fawad, M., Xu, X., & Hu, B. (2022). A framework for identification and classification of liver diseases based on machine learning algorithms. Frontiers in Oncology, 12, 1048348.

Dorina, K. (2012). Student performance prediction using data mining classification algorithms. International Journal of Computer Science and Management Research, 1.

Dritsas, E., & Trigka, M. (2023). Supervised machine learning models for liver disease risk prediction. Computers, 12(1), 19.

Erdal, H. I., & O. K. (2013). Advancing monthly streamflow prediction accuracy of CART models using ensemble learning paradigms. Journal of Hydrology, 477.

Fawcett, T., & Provost, F. J. (1996). Combining data mining and machine learning for effective user profiling. In Proceedings of KDD (pp. 8–13).

Ghosh, M., Raihan, M. M. S., Raihan, M., Akter, L., Bairagi, A. K., Alshamrani, S. S., & Masud, M. (2021). A comparative analysis of machine learning algorithms to predict liver disease. Intelligent Automation & Soft Computing, 30(3).

Giering, M. (2008). Retail sales prediction and item recommendations using customer demographics at store level. ACM SIGKDD Explorations Newsletter, 10(2), 84–89.

Gogi, V. J., & Vijayalakshmi, M. (2018). Prognosis of liver disease: Using machine learning algorithms. In 2018 International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE) (pp. 875–879). DOI: 10.1109/ICRIEECE44171.2018.9008482.

Gulia, A., Vohra, R., & Rani, P. (2014). Liver patient classification using intelligent techniques. International Journal of Computer Science and Information Technologies, 5(4), 5110–5115.

Gupta, K., Jiwani, N., Afreen, N., & Divyarani, D. (2022). Liver disease prediction using machine learning classification techniques. [Conference proceedings, pp. 221–226; specific conference not specified.]

Haloud, A. K. (2016). Selection of best decision tree algorithm for prediction and classification of student action. American International Journal of Research in Science, Technology, Engineering & Mathematics, 1, 26–32.

Hassoon, M., et al. (2017). Rule optimization of boosted C5.0 classification using a genetic algorithm for liver disease prediction. In 2017 International Conference on Computer and Applications (ICCA). IEEE.

Hogeweg, P. (2011). The roots of bioinformatics in theoretical biology. PLoS Computational Biology, 7(3), e1002021.

Jacob, J., Mathew, J. C., Mathew, J., & Issac, E. (2018). Diagnosis of liver disease using machine learning techniques. International Research Journal of Engineering and Technology (IRJET), 5(4).

Kamath, P. S., Wiesner, R. H., Malinchoc, M., Kremers, W., Therneau, T. M., Kosberg, C. L., D’Amico, G., Dickson, R. E., & Kim, R. W. (2001). A model to predict survival in patients with end-stage liver disease. Hepatology, 33(2), 464–470.

Karamouzis, S., & Vrettos, A. (2008). An artificial neural network for predicting student graduation outcomes. In Proceedings of the World Congress on Engineering and Computer Science, San Francisco, USA.

Karthik, S., Priyadarishini, A., Anuradha, J., & Tripathi, B. K. (2011). Classification and rule extraction using rough set for diagnosis of liver disease and its types. Advances in Applied Science Research, 2(3), 334–345.

Kefelegn, S., & Kamat, P. (2018). Prediction and analysis of liver disorder diseases by using data mining technique: Survey. International Journal of Pure and Applied Mathematics, 118, 765–770.

Kellis, M., et al. (2014). Defining functional DNA elements in the human genome. Proceedings of the National Academy of Sciences, 111(17), 6131–6138.

Khan, B., Naseem, R., Ali, M., Arshad, M., & Jan, N. (2019). Machine learning approaches for liver disease diagnosing. International Journal of Data Science and Advanced Analytics.

Khan, B., Shukla, P. K., Ahirwar, M. K., & Mishra, M. (2021). Strategic analysis in prediction of liver disease using different classification algorithms. In Handbook of Research on Disease Prediction Through Data Analytics and Machine Learning (pp. 437–449). IGI Global.

Kiruba, H. R., & Arasu, G. T. (2014). An intelligent agent based framework for liver disorder diagnosis using artificial intelligence techniques. Journal of Theoretical and Applied Information Technology, 69(1), 91–100.

Kotsiantis, S., Piarrekeas, C., & Pintelas, P. (2007). Predicting students’ performance in distance learning using machine learning techniques. Applied Artificial Intelligence, 18, 411–426.

Kovacic, Z. J. (2007). Early prediction of student success: Mining students enrolment data. Informing Science & IT Education Conference.

Kumar, S., & Katyal, S. (2018). Effective analysis and diagnosis of liver disorder by data mining. In 2018 International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 1047–1051). DOI: 10.1109/ICIRCA.2018.8596817.

Li, G., Luo, Y., Deng, W., Xu, X., Liu, A., & Song, E. (2008). Computer-aided diagnosis of fatty liver ultrasonic images based on support vector machine. In 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS 2008) (pp. 4768–4771).

Liu, J., Liu, B., Liu, Y., Chen, H., Feng, L., Xiong, H., & Huang, Y. (2017). Personalized air travel prediction: A multi-factor perspective. ACM Transactions on Intelligent Systems and Technology (TIST), 9, 1–26.

Ma, H. Y., Zhou, Z., Wu, S., Wan, Y. L., & Tsui, P. H. (2016). A computer-aided diagnosis scheme for detection of fatty liver in vivo based on ultrasound kurtosis imaging. Journal of Medical Systems, 40(1), 1–9.

Marques-Vera, C., & Cano, A. (2013). Predicting student failure at school using genetic programming. Applied Intelligence, 38, 315–330.

Midhila, M., Krishnan, K. R., & Sudhakar, R. (2017). A study of the phases of classification of liver diseases from ultrasound images and gray level difference weights based segmentation. In 2017 International Conference on Communication and Signal Processing (ICCSP) (pp. 1582–1587). DOI: 10.1109/ICCSP.2017.8286655.

Minei-Bidgoli, B., Kashy, D., Kortemeyer, G., & Punch, W. F. (2003). Predicting student performance: An application of data mining methods with the educational web-based system LON-CAPA. In 33rd ASEE/IEEE Frontiers in Education Conference.

Mumbower, S., Garrow, L. A., & Higgins, M. J. (2014). Estimating flight-level price elasticities using online airline data. Transportation Research Part A: Policy and Practice, 66, 196–212.

Musleh, M. M., Alajrami, E., Khalil, A. J., Abu-Nasser, B. S., Barhoom, A. M., & Naser, S. A. (2019). Predicting liver patients using artificial neural network. International Journal of Academic Information Systems Research (IJAISR), 3(10).

Nahar, N., & Ara, F. (2018). Liver disease prediction by using different decision tree techniques. International Journal of Data Mining & Knowledge Management Process, 8(2), 01–09.

Nguyen, N. N., Janeck, P., & Haddawy, P. (2007). A comparative analysis of techniques for predicting academic performance. In 37th ASEE/IEEE Frontiers in Education Conference.

Olaniyi, E. O., & Aadnan, K. (2015). Liver disease diagnosis based on neural networks. In Advances in Computational Intelligence, Proceedings of the 16th International Conference on Neural Networks (NN '15), November 7–9, 2015.

Osmanbegovic, E., & Suljic, M. (2012). Data mining approach for predicting student performance. Journal of Economics and Business, 10(1).

Owjimehr, M., Danyali, H., Helfroush, M. S., & Shakibafard, A. (2017). Staging of fatty liver diseases based on hierarchical classification and feature fusion for back-scan-converted ultrasound images. Ultrasonic Imaging, 39(2), 79–95.

Pahariya, J., Vohra, R., Makhijani, J., & Patsariya, S. (2014). Liver patient classification using intelligence techniques. International Journal of Advanced Research in Computer Science and Software Engineering, 4(2), 295–299.

Pal, A. K., & Pal, S. (2013). Data mining techniques in EDM for predicting the performance of students. International Journal of Computer and Information Technology, 2(6).

Panjwani, M., Ramrakhiani, R., Jumnani, H., Zanwar, K., & Hande, R. (2020). Sales prediction system using machine learning. EasyChair Preprint No. 3243.

Patnaik, R. K., Lin, Y.-C., Agarwal, A., Ho, M.-C., & Yeh, J. A. (2022). A pilot study for the prediction of liver function related scores using breath biomarkers and machine learning. Scientific Reports, 12(1), 2032.

Priya, M. B., Juliet, P. L., & Tamilselvi, P. (2018). Performance analysis of liver disease prediction using machine learning algorithms. International Research Journal of Engineering and Technology, 5(1), 206–211.

Ragg, T., Menzel, W., Baum, W., & Wigbers, M. (2002). Bayesian learning for sales rate prediction for thousands of retailers. Neurocomputing, 43(1–4), 127–144.

Rahman, A., Shamrat, F. J. M., Tasnim, Z., Roy, J., & Hossain, S. A. (2019). A comparative study on liver disease prediction using supervised machine learning algorithms. International Journal of Scientific & Technology Research, 8(11), 419–422.

Rajeswari, P., & Reena, G. S. (2010). Analysis of liver disorder using data mining algorithm. Global Journal of Computer Science and Technology.

Ramana, B. V., Babu, M. S. P., & Venkateswarlu, N. B. (2011). A critical study of selected classification algorithms for liver disease diagnosis. International Journal of Database Management Systems (IJDMS), 3(2).

Ramana, B. V., Babu, M. S. P., & Venkateswarlu, N. B. (2012). Liver classification using modified rotation forest. International Journal of Engineering Research and Development, 1(6), 17–24.

Ramaswami, M., & Bhaskaran, R. (2010). A CHAID based performance prediction model in educational data mining. International Journal of Computer Science Issues, 7(1), 1.