Chronic Liver Disease Classification Using Machine Learning Techniques

Israel Merdasa Umeta

A Thesis Submitted to

The Department of Computer Science and Engineering

School of Electrical Engineering and Computing

Presented in Partial Fulfillment of the Requirement for the Degree of Masters

in Computer Science and Engineering

Office of Graduate Studies

Adama Science and Technology University

February 2022

Adama, Ethiopia

Chronic Liver Disease Classification Using Machine Learning Techniques

Israel Merdasa Umeta

Advisor: Tilahun Melak (Ph.D)

A Thesis Submitted To

The Department of Computer Science and Engineering

School of Electrical Engineering and Computing

Presented in Partial Fulfillment of the Requirement for the Degree of Master's

in Computer Science and Engineering

Office of Graduate Studies

Adama Science and Technology University

February 2022

Adama, Ethiopia

Declaration

I hereby declare that this MSc Thesis entitled “Chronic Liver Disease Classification Using Machine Learning Techniques “is my original work. That is, it has not been submitted for the award of any academic degree, diploma or certificate in any other university. All sources of materials that are used for this thesis have been duly acknowledged through citation

Name: Israel Merdasa

Signature: _____

Date: _____

Recommendation

I/we, the advisor(s) of this thesis, hereby certify that I have read the revised version of the thesis entitled “Chronic Liver Disease Classification Using Machine Learning Techniques “prepared under my guidance by Israel Merdasa submitted in partial fulfillment of the requirements for the degree of Masters of Science in Computer Science and Engineering. Therefore, I recommend the submission of the revised version of the thesis to the department following the applicable procedures.

_____	_____	_____
Advisor	Signature	Date

Approval Page of M.Sc. Thesis

I/we, the advisor(s) of the thesis entitled "Chronic Liver Disease Classification Using Machine Learning Techniques" and developed by Israel Merdasa hereby certify that the recommendation and suggestions made by the board of examiners are appropriately incorporated into the final version of the thesis.

Advisor	Signature	Date
Co- advisor	Signature	Date

We, the undersigned, members of the Board of Examiners of the final open defense by Israel Merdasa have read and evaluated his thesis entitled "Chronic Liver Disease Classification Using Machine Learning Techniques" and examined the candidate. This is, therefore, to certify that the thesis has been accepted in partial fulfillment of the requirement of the Degree of Masters of Science in Computer Science and Engineering.

Chairperson	Signature	Date
Internal Examiner	Signature	Date
External Examiner	Signature	Date

Final approval and acceptance of the thesis is contingent upon submission of its final copy to the Office of Postgraduate Studies (OPGS) through the Department Graduate Council (DGC) and School Graduate Committee (SGC).

Department Head	Signature	Date
School Dean	Signature	Date
Office of Postgraduate Studies, Dean	Signature	Date

ACKNOWLEDGMENTS

First, I would like to thank God for all his blessing from the beginning to the end of this research. I am extremely grateful to my advisor Dr.Tilahun Melak for his friendly approach, invaluable advice, critical comments, and commitment to guide, and continuous support that greatly improved this thesis work.

I would also like to acknowledge the members of the Smart Software SIG of the CSE department. I am grateful to the SIG supervisor Dr. Mesifn Abebe for his continuous commitment to guide and support this research from its inception to completion. I would like to thank all my classmates who are working on their master's thesis for their cherished time spent together, ideas, and support until the end of my thesis.

I would also like to thank the Staff of Nekemte Referal Hospital (NRH) for helping, allowing, and providing me with valuable information to collect the data. My appreciation also goes out to my family and friends for their encouragement and support all through my study. I would also like to express my gratitude and appreciation to Duresa Tamirat (PhD. Candidate at ASTU) for his support and continuous encouragement throughout my study.

Finally yet importantly, my thanks go to my lovely brother Amanuel Merdasa, who would have not been possible without this.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ALP	Alkaline –Phosphatase
ALT	Alamine -aminotransferase
ANN	Artificial Neural Network
AST	Aspartate-Aminotransferase
BMC	Biomedical and Chemical
CHIRP	Composite Hypercube on Iterated Random Projection
CLD	Chronic Liver Disease
CPU	Central Processing Unit
CSV	Comma Separated Value
CV	Cross Validation
DB	Direct Bilirubin
DL	Deep Learning
ESLD	End Stage Liver Disease
FN	False Negative
FP	False Positive
GBD	Global Burden Disease
GH	Giga Hertz
GUI	Graphical User Interface
HBV	Hepatitis B Virus
HCV	Hepatitis C Virus
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
IDE	Integrated Development Environment
KNN	K-Nearest Neighbour
LR	Logistic Regression
ML	Machine Learning

MST	Minimum Spanning Tree
NAFLD	Non-Alcoholic Fatty Liver Disease
NASH	Non-Alcoholic Steato Hepatitis
NB	Naïve Bayes
NCD	Non-Communicable Disease
NRH	Nekemte Referal Hospital
OVR	One-Vs- Rest
RF	Random Forest
SVC	Support Vector Classifier
SVM	Support Vector Machine
TB	Direct Bilirubin
TN	True Negative
TP	True Positive
WHO	World Health Organization

ABSTRACT

The chronic liver disease provides significant challenges to the global healthcare system, as it consumes the majority of healthcare budget, particularly in low-income nations. Early classifying and prevention based on disease severity is among the most popular significant issues facing the health sector, especially in developing countries such as Ethiopia. Machine learning has an important role in analyzing a huge amount of medical information and addressing complex problems for early disease classification. The goal of this research was to use machine-learning algorithms to classify chronic liver disease as not chronic liver disease or chronic liver disease, depending on whether the disease was present or not. In this study, the experimental design have been used with an experimental approach to determine the most appropriate performance model. The data for this study purpose was collected from Nekemte Referal Hospital. The three algorithms employed in the study were support Vector Machine, Logistic Regression, and Nave Bayes. Evaluation of the models was done using 10-fold cross-validation and classification performance in order to compare the models. The performance of the given classification techniques was evaluated using accuracy, precision, recall, F1-score, sensitivity, and specificity. Logistic Regression (LR), Support Vector Machine (SVM), and Naive Bayes have accuracy rates of 84%, 92%, and 83%, respectively. The study findings show that, Support Vector Machine record better performance compared to Logistic Regression and Naïve Bayes models based on accuracy and Fl-score.

Key Words: *Machine learning, Chronic Liver Disease, Logistic Regression, Support Vector Machine, Naïve Bayes.*

CHAPTER ONE

INTRODUCTION

1.1. Background of the Study

Chronic diseases, also known as non-communicable diseases, are long-term illnesses caused by a mix of genetic, physiological, environmental, and behavioral factors(WHO, 2020). Chronic liver disease (CLD) is also a non-communicable disease in which the liver tissue undergoes gradual damage and restoration, resulting in liver cirrhosis. Chronic liver disease is a long-term condition in which the parenchyma of the liver is continually damaged and replaced with fibrous tissue, resulting in cirrhosis(Blundell et al., n.d.). Both inside and outside the body, the liver is the biggest organ. It provides support to practically all of the body's organs and conducts a range of important functions, including filtering pollutants and dangerous synthetic compounds. The liver is involved in a number of ways bodily tasks, including the synthesis of proteins, blood control, cholesterol, sugar, and iron metabolism. It is necessary for living and serves a variety of activities, including eliminating toxins in the body. Hepatitis, addiction, and other viruses create chronic diseases. Long-term infection with the viral hepatitis B virus (HBV), viral hepatitis C virus (HCV), or another hepatitis virus causes chronic liver disease. As the prevalence of chronic liver disease rises over the world, it places a considerable burden on low-income countries where detection, prevention, and treatment rates are low(Bollyky et al., 2017).

According to the World Health Organization's Global Hepatitis Report (WHOGHR) 2019, 1.1 million deaths from cases of viral hepatitis have been recorded. As a result, chronic liver disease is one of the world's most serious health issues. Every year, around 2 million individuals die from chronic liver disease, 1 million from cirrhosis complications, and 1 million from hepatocellular carcinoma and viral hepatitis around the world (Sontakke et al., 2017). According to the Global Burden of Disease Study (GBD), chronic liver disease was responsible for 2.36% of fatalities in 2017. In the United States, the Centers for Disease Control and Prevention reported a 1.5% mortality rate due to chronic liver disease in 2015. Chronic liver disease is a major public health concern in Ethiopia today, affecting a higher number of people regardless of their age, gender, or religious affiliation. Chronic liver disease mortality in Ethiopia reached 16,580 in 2018, accounting for 2.71 % of all deaths, according to the latest WHO data. With an age-adjusted Death Rate of 30.87 per 100,000 population, Ethiopia is ranked number 33 in the world.

This condition is due to a scarcity of proper diet and physical activity. Many factors can damage the liver, resulting in chronic liver disease. Obesity, undetected hepatitis, alcohol misuse that causes neurological dysfunction, hematemesis, liver failure, jaundice, and hepatic encephalopathy are the conditions that can induce neurological dysfunction. Chronic liver disease is not known to exist in Ethiopia, but it is widely believed to exist. Similarly, the etiology has been used to estimate the percentage burden of chronic liver disease, but no data for Ethiopia is currently available. Chronic HBV or HCV infection causes more than 60% of chronic liver damage. Hepatitis B surface antigen (HBsAg) prevalence is predicted to be 6.0% in Ethiopia, and HCV antibody (antiHCV) prevalence is expected to be 3.1%(Krishnan, 2019). Machine learning is employed in the creation of information easier to understand for decision-makers, and it plays an important role in disease classification based on high-performance medical history data (Kohli & Regression, 2015). This research is focused on using machine-learning techniques to classify chronic liver disease based on datasets collected from the Nekemte Referral Hospital in Ethiopia.

1.2. Motivation of the Study

According to various health organizations' reports and statistical analyses, the number of persons suffering with CLD is increasing from time to time around the world. Chronic Liver disease is becoming a severe problem, especially in developing nations like Ethiopia. Many researchers have conducted different research on how to manage the progress of chronic liver disease locally and globally. Most globally conducted research on chronic liver disease classification focuses on the identification of the presence or absence of the disease but not on the identification of the severity of the disease. In Ethiopia, different statistical analysis research was conducted on the prevalence and the quality of life of the patient with chronic liver disease. Furthermore, in Ethiopia, the treatment of liver disease is given in only specified hospitals to which thousands of people run to get the treatment and there is not enough nephrologists and computer-aided diagnosis to make the fast diagnosis.

With this concern, nowadays, computer technology and machine learning techniques are highly in use to develop software to assist doctors in deciding the status of chronic liver disease in the preliminary stage. One paper was conducted by Siraj Mohammed and Tibebe Beshah (Cheemerla & Balakrishnan, 2021) in Ethiopia on the treatment and diagnoses of chronic liver disease using KBS and the machine-learning technique. However, the focus of the paper was only on the KBS.

Therefore, we were inspired to work on the CLD classification using machine learning techniques because machine learning can analyze medical data and classify the severity of the disease correctly by simply inserting common clinical predictors. Using machine-learning techniques assists and enables the experts to make fast decisions and take necessary action to treat patients and give an appropriate recommendation based on the severity of the disease. Preparing the new dataset from locally collected patient history data, the study will contribute to the development of a better CLD Classification system to save the time and cost of treatment with the help of machine learning models.

1.3. Statement of the Problem

Noncommunicable diseases (NCDs) are the leading cause of death and disability worldwide, regardless of geographic location, population size, social class, or economic development. It increases day by day and is difficult to find easily in the early stages. For many years, non-communicable diseases were not been considered as priority problems in many low-level income countries including Ethiopia. Chronic liver disease can progress to liver failure if it does not detect and treated early which consumes a huge budget and is even difficult for developing countries like Ethiopia to get treatment locally. In developing countries like Ethiopia, the patients get the treatment after a serious case due to lack of medical resources, specialist doctors, and low income for treatment(Tesfaye et al., 2021). Chronic liver disease treatment is not widely available in most of Sub-Saharan Africa, including Ethiopia. For many years, there has been little research into the treatment and diagnosis of chronic liver disease in Ethiopia, and there is no national policy for the prevention and treatment of chronic liver disease patients. In Ethiopia, most of the entire hospital system is done manually, and in some hospitals, the patient's medical history is recorded on paper, making it difficult to get the information. Additionally, there is a widespread lack of awareness among people and health centers about the treatment of chronic liver disease, as well as awareness of technologies like machine learning, data mining, and data analytics for early disease detection, and even the lack of appropriate local hospitals. Furthermore, many people are unaware of the disease and are only aware that they have it when their livers become chronic and have an impact on their lives. They believe that all of their symptoms and agony are due to common cold and that they will be OK if they take some medications without getting identified. That is bad thinking and several factors encourage this bad attitude such as the traveling distance to the clinic or hospital which takes a lot of time and waiting time especially in the governmental hospital to get treatment.

Thus the early detection enables the expert to diagnose the patients fast to reduce waiting time and can slow the progress and even stop the CLD and helpful in the management of other risk factors. Because clinical data is so complicated, today's modern healthcare system has the challenge of gathering, evaluating, and applying a vast amount of knowledge needed to overcome difficult clinical problems. A major challenge facing healthcare institutions (hospitals, medical centers) is the provision of quality services at affordable costs because the quality service implies the timely diagnosis of patients and the timely management of treatments that are efficient and effective. Using the machine learning techniques will solve the challenge of the healthcare system on disease classification by training the machine with appropriate data. Thus, the proposed model could be beneficial for physicians; health professionals determine the severity of chronic liver disease (Online et al., 2021).

Therefore, the contribution of this paper is to prepare a dataset containing instances with relevant features to classify chronic liver disease using machine-learning algorithms. Thus, this system could help physicians, and health professionals to detect the disease and give fast treatment for the patients in a short time. The main aspect of this work is to improve the early treatment and diagnosis of chronic diseases in people in low-income and developing countries. As a result of our findings, machine-learning approaches could be used to classify chronic liver disease (Aravind et al., 2020).

This study addresses the problem by answering the following question:

- What are the attributes to consider designing effective chronic liver disease datasets in applying machine-learning models?
- Which machine-learning classifier is best appropriate for chronic liver disease classification?

1.4. Objective

1.4.1. General Objective

The general objective of this study is to develop Chronic Liver Disease classification using machine-learning techniques.

1.4.2. Specific Objective

- To review the literature on the concept of machine learning algorithms to classify chronic liver disease.
- To collect different symptoms and laboratory test results of CLD and non-CLD patients and prepare the dataset for training and test.
- To increase the accuracy of current diagnostic approaches for the classification of chronic liver disease.
- To find the best classification models for CLD.

1.5. Scope and Limitation of the Study

1.5.1. Scope

There are several approaches to developing a disease classification system. This study is focused on applying machine-learning algorithms to classify the disease as cld or notcld that enables experts can provide appropriate treatments and recommendations based on datasets developed from locally collected medical history data from Nekemte Referral Hospitals. Binary class records are also used to determine whether a disease is present or not. To classify chronic liver disease, researchers employed a supervised machine learning approach with 9 features from the original dataset.

1.5.2. Limitation

The research comes across different limitations on a different phase of the research processes. The most difficult thing in Ethiopia especially in the medical area, it is very difficult to find the organized data in electronic format and it is time-consuming to transform the data from manually recorded to electronic data. Data were only collected from one hospital. However, building the model with a dataset collected from various medical centers and external validation could have better performance and be more reliable. The researcher found only one study on chronic liver disease classification. It does not consider other chronic diseases

1.6. Applications of Results

The main beneficiaries of the proposed work will be the health sectors. Mainly physicians will benefited from this disease classification system. Additionally the patients are benefit from the system because the classification model reduce the number of attributes to be use in chronic liver disease classification and save the cost for treatment.

1.7. Organizations of the Thesis

This thesis is organized into seven chapters.

Chapter One Introduction: describes the introduction and covers the motivation, the statement of problems, the objective of the study, the scope, limitation, and application of the study, and the organization of the study.

Chapter Two Literature Review: states the literature related to this study, reviews background information on Chronic Liver disease and algorithms used in CLD and finally related works are discussed.

Chapter Three Methodology: the third chapter states, research methodology used in this study, methods used to build the dataset, methods used to develop chronic Liver disease classification system which are data preprocessing, and machine learning classifier algorithm's, and finally, it states the evaluation methods.

Chapter Four Proposed Solution: discusses the proposed solution for chronic Liver disease classification, which is the proposed data preprocessing, machine learning classifiers and evaluation methods.

Chapter Five Implementation: presents the implementation and experimentation of the proposed solution. Including working environment, dataset description, also the implementation of preprocessing, model implementations, and evaluation models.

Chapter Six Results and Discussion: discusses the result of the dataset class distribution, binary class, and four-class models and discusses the major results obtained by comparing the models based on their performance on both binary classes.

Chapter Seven Conclusion: describes conclusion and forwards recommendation and feature works for further investigation of this research study.

CHAPTER TWO

LITERATURE REVIEW AND RELATED WORKS

In this chapter, the relevant literature related to the main objective of the paper is reviewed to clearly state and furtherly explore the research problem by providing a technical review on chronic liver disease and existing method to detect chronic liver disease. Started by stating the background chronic Liver disease, chronic Liver stages, and overview of machine learning, machine learning and health center, classification algorithm, and previously worked studies to solve detection problem chronic Liver disease.

2.1. Chronic Liver Disease and Its Risk Factors

Chronic liver disease is a progressive disorder in which the liver's function breaks down over time. The liver is responsible for the production of clotting factors and other proteins, as well as the detoxification of harmful substances and the elimination of bile. Chronic liver disease causes fibrosis and cirrhosis due to a continuous cycle of inflammation, destruction, and regeneration of the liver parenchyma. Toxins, long-term alcohol abuse, viral infections, and autoimmune illnesses, hereditary and metabolic abnormalities, cause chronic liver disease. Chronic liver disease is the leading cause of death worldwide, particularly in developing nations. In developed countries, the most common chronic liver disease is alcoholic liver disease, followed by chronic viral hepatitis (hepatitis B and C), non-alcoholic fatty liver disease (NAFLD), and hemochromatosis(Liu et al., 2018)(Durai, n.d.). The goal of this work is to use machine-learning approaches to produce chronic liver disease classification(Banu Priya et al., 2018).

Chronic liver disease is a major medical and public health issue that results in significant morbidity and mortality all over the world. Chronic liver disease is influenced by the prevalence of causes such as hepatitis B virus (HBV) and hepatitis C virus (HCV) infections, alcohol consumption, and viral hepatitis infections such as diabetes. According to research published in the American Journal of Gastroenterology, obesity is the most common risk factor for chronic liver disease. Obesity is one of the world's most severe disorders, with an estimated 2.1 billion overweight persons globally in 2013. Chronic liver disease was determined to be responsible for 2% of all deaths, according to a recent estimate published in BMC Medicine based on the Global Burden of Study (GBD) research(Byass, 2014). In today's world, chronic liver disease is one of the most common causes of death.

Chronic liver disease risk factors differ from one community to the next. In developed countries, alcohol, chronic hepatitis B virus, chronic hepatitis C virus, and nonalcoholic steatohepatitis (NASH) are risk factors, however, in underdeveloped countries like Ethiopia, alcohol, type B hepatitis, and hepatitis C are the primary causes of cirrhosis(Banait et al., 2021).

2.1.1. Worldwide prevalence of Chronic Liver Disease

Chronic liver disease is a major public health issue that affects millions of people around the world. The prevalence of viral hepatitis is increasing in most developed nations because of recent advances in disease prevention, detection, and treatment. Chronic diseases account for over 60% of global mortality, with chronic liver disease accounting for nearly 2 million deaths every year. More than 2 million individuals died in 2010 as a result of major chronic liver disease, such as acute hepatitis, cirrhosis, and liver cancer, according to the Global Burden of Disease research.

2.1.2. Prevalence of Chronic Liver Disease in Ethiopia

Chronic diseases cause around 60% of all deaths worldwide, with chronic liver disease killing nearly 2 million people each year. Ethiopia is a low-income country in East Africa with a population of approximately 100 million people(Magnus & Orlien, 2019). The frequency of chronic liver disease in Ethiopia is unclear, however, it is considered significant. Similarly, while it has been reported that persistent HBV or HCV infections are responsible for more than 60% of chronic liver disease cases, the proportional burden of chronic liver disease by etiology remains unknown. In Ethiopia, 2.1% of alcohol-related problems are less common than in both regional (WHO Africa: 3.3%) and global estimations (4.1%).

There are no current representative nationwide data on the prevalence of alcoholic liver disease in Ethiopia as far as we know, however, it is thought to be below the global average(Jeong et al., 2017). Although there are no prevalence studies on NAFLD in Ethiopia, it is anticipated to be low because Ethiopia has one of the lowest prevalence rates of BMI 25.0 kg/m2 in the world. Chronic liver disease is a worldwide health problem with a high rate of morbidity and mortality. In low and middle-income countries, viral hepatitis is the major cause of liver disease, and it is also widespread in high-income countries, but primarily as a side effect of drug abuse. In Ethiopia, viral hepatitis B and C are widespread, but chronic liver disease has received little attention

2.1.3. Chronic liver disease stages

Chronic liver disease is a leading cause of death worldwide, particularly in developing countries. When scar tissue in the liver replaces good tissue, the chronic liver disease develops. The chronic liver disease must be staged since it is progressive, asymptomatic for the most part, and possibly fatal. Chronic liver disease is divided into four stages.

Stage1: Inflammation

Due to ongoing damage to liver cells produced by a number of chemicals and illnesses, the liver is enlarged at this stage. At this time, the disease may be treated. Patients with non-alcoholic fatty liver may show no symptoms at all.

Stage 2: Scarring

If the inflammation in stage 1 is not treated, the liver tissues begin to scar. Scarred liver tissue starts to replace good liver tissue. As toxins, fat, and damaged tissue accumulate in the liver, fibrosis can impair its architecture and disrupt blood flow inside it. Even at this stage, with the correct medication and lifestyle changes, the sickness can be reversed to some extent.

Stage 3: Cirrhosis

The liver will not be able to recover on its own at this point, because the scarring is complete. For such an individual, a liver transplant is the only alternative. The area of the abdomen where the liver is located is in a lot of pain, discomfort, and tenderness. The person loses their appetite, and anything they consume produces bloating in the abdomen due to fluid build-up. He or she also has jaundice, which causes yellowing of the eyes and skin. The accumulation of bile pigments in the skin frequently causes itching.

Stage 4: End-stage liver disease

As cirrhosis progresses, the liver begins to fail. Liver failure can be divided into two categories. Acute liver failure occurs quickly, usually within 48 to 72 hours, and can be caused by a number of factors other than alcohol. Alcohol abuse is a leading cause of chronic liver disease, which can take years to develop. Symptoms include diarrhea, nausea, vomiting, weight loss, and a loss of appetite. The person's mental or cognitive health is also damaged, and he or she is frequently confused. ESLD is almost always deadly. The victim can be put on life support until a matching liver is found.

2.2. Machine learning

Machine learning is an artificial intelligence discipline in which a computer program learns from its prior experiences and performances(Technology, 2017). The majority of enterprises around the world, allowing them to save time and money while increasing revenues, now uses a computerized system. It is one of the most essential and intriguing ways for solving complex challenges with large amounts of data in today's studies. Machine learning aids in the discovery of hidden connections and linkages in huge datasets, resulting in more precise classification and model creation. Its purpose is to teach the machine what to do and then have it do it without having to write it explicitly. Machine learning can be classified as supervised, unsupervised, semi-supervised, or reinforcement learning, according to experts.

2.2.1. Supervised learning

One of the most popular methods for extracting knowledge from databases containing previously known training samples is supervised learning. A target or dependent feature is selected from a group of independent features in this algorithm. Data that is labeled or categorized is employed in the supervised learning method. The learned rule is then applied to unobserved data with uncertain outputs to categorize them. Classification and regression are the two most common supervised machine-learning problems. Supervised machine learning methods include Naive Bayes, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Logistic Regression (LR), and Decision Tree (DT).

2.2.2. Unsupervised learning

Unsupervised learning is a method of working with data that has not been labeled. This method is a strong tool for evaluating data and discovering patterns and trends in the absence of a known label for the instances. It is a learning algorithm that may be used to divide datasets into categories. K-means are the best examples of unsupervised learning.

2.2.3. Semi-supervised Learning

Semi-supervised learning combines supervised and unsupervised machine learning techniques. The classified data are mixed in this type of method. This technique seeks to learn a model that predicts future test data targets better than a model created just from data with a known target.

2.2.4. Reinforcement Learning

This algorithm tries to exploit data acquired from interactions with the environment, and it continuously trains itself through trial and error and feedback mechanisms. This algorithm learns from experiences and strives to use the most up-to-date information available to make appropriate conclusions. The Markov Decision Process is used to demonstrate reinforcement learning.

2.3. Machine Learning and Health Sector

The healthcare industry has a long history of being an early adopter of new technology and enjoying significant benefits as a result. Machine learning is being applied in a number of health-related sectors, including drug research, data and record management, and chronic disease therapy. For a country to take care of its people's lives, the health industry is a critical sector. For developing countries like Ethiopia, where the majority of the population is rural and subsistence farmers with limited access to safe drinking water, housing, sanitation, food, and health care, good health care is critical. One crucial aspect of improving healthcare is to place a greater emphasis on disease prevention. Not every disease can be prevented, but in many cases, early detection can lead to better health outcomes and cheaper treatment costs. Early disease classification is one method of preventing healthcare complications. The diagnosis that occurs early in the disease process has the potential to save lives. Patients' lives can be extended significantly if they are diagnosed early(Jacob et al., 2018). Machine learning can be used in healthcare for a variety of activities such as disease detection and diagnosis, drug discovery and production, medical imaging diagnosis, personalized medication, and classification.

2.3.1. Machine Learning for Disease Classification

With the rapid advancement of technology and data, the healthcare domain is one of the most significant study fields in the contemporary era. Machine learning is an emerging approach that helps in the classification and diagnosis of disease. Machine learning plays a crucial role in disease classification based on patient history data. The machine learning technique showed success in the healthcare system in the classification and diagnosis of numerous critical diseases. Machine learning in the healthcare system enables the practitioners to process huge and complex medical datasets and then analyze them into clinical insights. A disease's classification is based on the patient's symptoms and medical history. Machine

learning technology provides a good platform in the medical industry for rapidly resolving healthcare challenges(Pingale et al., 2019).

2.5. Related Works

This section discusses the review of related research papers to chronic liver disease classification to clearly understand general techniques, methods, and results of current studies. Many researchers have been researching the system that can classify the diseases and still the health sectors are left with huge researchable areas that need the intervention of machine learning suitable for the classification of disease. Different related works have been conducted locally and globally by different researchers in the health domain using machine-learning techniques have reviewed.

Dr. S. Vijayarani and Mr. S. Dhayanand *et al* (Vijayarani, 2015) proposed classification models such as Support Vector Machine and Nave Bayes to classify chronic liver disease based on performance evaluation and execution time. The researchers are used a dataset 466 from the UCI repository with 10 features. Researchers have used and compared two classifiers 'algorithm. Finally, they arrived to the conclusion that SVM is a better classifier for the classification of chronic liver disease.

Himanshi Bansal, Kapil Sharma, and Goldendeep Kaur *et al*(Bansal et al., 2020) proposed a hybrid approach, which uses the concept of clustering and classification for chronic liver disease. The researcher are used a dataset 583 records of patients taken from UCI repository with 11 features. The algorithms used in this work are K-means clustering, Random Forest classifier. Researchers is compared with existing classifiers like KNN, LR, etc using the confusion matrix parameters such as precision, recall, f1-score and accuracy. Finally, hybrid approach shows better results than other algorithms for the classification chronic liver disease.

Shapla Rani Ghosh and Sajjad Waheed *et al* (Ghosh & Waheed, 2017) proposed an Analysis of classification algorithms for liver disease diagnosis using Naïve Bayes, Bagging , KStar, Logistic Regression and Rep tree were used to evaluate the accuracy, precision and specificity. Researchers are used two datasets AP and UCI were considered to find out best algorithm. The dataset contains 583 instances with 10 attributes and the second dataset contains 345 instances with 6 attributes. For the purpose of experimentation and analysis researches use weka (version 3.6.10) data mining software. Finally, KStar had maximum accuracy for classification of liver disease.

Binish Khan, Piyus Kumar Shukla, and Manish Kumar Ahirwar *et al* (Binish Khan et al., 2019) Proposed Strategic Analysis in Liver Disease Classification using machine learning techniques. To determine the optimum strategy, they compare various classification algorithms such as Random Forest, Logistic Regression, and Separation algorithm. Finally, Random Forest has the highest classification accuracy for liver disease classification.

Thirunavukkarasu K, Ajay S. Singh, Md Irfan, and Abhishek Chowdhury *et.al* (Thirunavukkarasu et al., 2018) proposed classification of liver disease using machine-learning algorithms such as, Logistic Regression (LR), K-Nearest Neighbour (KNN), and Support Vector Machine (SVM). The researchers are used a dataset taken from Indian Liver Patient Dataset (ILPD) contains 567 records with 10 attributes. Researchers also compare accuracy score and a confusion matrix. Finally, the best model for classifying liver disease is logistic regression.

Sivakumar D Manjunath Varchagall Ambika L GUsha S *et.al* (Sivakumar et al., 2019) proposed classification models to recognize and analyze the human life quality attributes in forecasting the chronic liver disease with machine learning techniques such as, K-means clustering calculation and C4.5 decision tree approaches with distinctive error measures ,such as RMSE, MAE and Kappa measurement esteems. The researchers used the proportion of 80:20 for preparing and testing the chronic liver disease forecast framework.

Mehta Banu H *et.al* (Mehtaj Banu, 2019) worked on "the classification of liver disease using machine-learning techniques" such as KNN and SVM. Researchers used a dataset taken from UCI repository contain 416 instances with 6 attributes. Finally, support vector machine records highest accuracy to classifying liver disease.

Hoon Jin , Seoungcheon Kim, and Jinhong Kim *et al*(Jin et al., 2014) worked on "Decision Factors on Effective Liver Patient Data Prediction" such as, Naïve Bayes, Decision Tree, Multi-Layer Perceptron and KNN. The researchers are used a dataset contain 679 instances with 10 features. It also compare the performance evaluation metrics such as, precision, recall, sensitivity, specificity. Finally, Naïve Bayes took precedence over other algorithms in the performance of prediction test as considering the algorithmic characteristics to liver patient dataset.

Table 2. 1 Some summary of Related works

No	Title	Methodology and Accuracy
1	Chronic liver disease classification using SVM and Nave Bayes [Vijayarani, 2015]	SVM and Naive Bayes with an accuracy of 79.66% and 61.28% respectively.
2	A hybrid approach for the classification of the liver disorder using data mining [Bansal et al., 2020]	Random Forest, KNN, LR and Hybrid approach with an accuracy of 82.06%, 72.05%, 72.33% and 85% respectively
3	Analysis of classification algorithms for liver disease diagnosis (Ghosh & Waheed, 2017)	Naïve Bayes, and Logistic Regression with an accuracy of 69%, 75.4% respectively.
4	Strategic Analysis in classification of Liver Disease Using Different machine learning algorithms [Binish Khan et al]	Random Forest and Logistic Regression with an accuracy of 85% and 83.41%
5	classification of liver disease using machine learning algorithms (Thirunavukkarasu et al., 2018)	K-NN, Logistic regression and Support Vector Machine (SVM) with an accuracy of 73.97% 73.97% and 71.97% respectively.
6	Chronic liver disease classification analysis based on the impact of life quality attributes (Sivakumar et al., 2019)	C4.5 and K-means clustering with an accuracy of 84%, 36% and 83.7% respectively**.**

CHAPTER THREE

METHODOLOGY

This chapter discusses the methodology and the process of developing machine-learning techniques to classify chronic liver disease. Nowadays, the application of machine learning is quickly increasing in various sectors. Health sectors are among those sectors, which are using machine-learning technology extensively to satisfy the customer's needs. Because the health sectors have a huge amount of data, which is complex for the human for analyze, but machine learning makes it easy to make a pattern and analysis of the data. Different methods, procedures, and techniques have to be followed and used in this research. The justification of the methods and procedures have been made in different phases. First, the data collection method and techniques are discussed. The second phases explains the preprocessing method, which is a very important step in machine learning model building. The third phase explains and justifies the supervised machine learning algorithms used in developing a chronic liver disease classification. Finally, the performance evaluation method used in this research study is explained.

3.1. Data source

The data to conduct this research was collected from the patient history data from Nekemte Referral Hospital (NRH). The major public hospital in Nekemte is the Nekemte Referral Hospital. Nekemte Referral Hospital is the main government-run healthcare facility in Nekemte, which is 331 kilometers from Addis Ababa. The sole referral hospital in the East Wollega zone, Nekemte Referral Hospital provides a variety of health services to over 1,756,952 people. Swedish missionaries founded the Nekemte Referral Hospital in 1932. It employs 8 specialists, 11 general practitioners, 71 nurses, 9 health officers, 8 laboratory technicians, 7 pharmacy technicians, 2 sanitarians, and 78 administrative employees. There are 178 beds at the hospital. It opens up for the services and has a good opportunity for researchers and policymakers for program monitoring and evaluation.

3.2. Data collection

The data for this study was secondary and collected from the patient history data without direct communication with the patient. The data collected and used in this study, included patients' records of chronic liver disease, who were admitted to attend their treatment or died during a period 2016 to 2020, and some of them were obtained from the same patient history data at different times. To collect data different features to classify the presence of chronic liver disease are identified by asking the domain experts and considering online available features of the chronic liver disease dataset.

3.3. Preprocessing

In our research, preprocessing is the most important step in creating a classification model. Preprocessing is the process of transforming noisy, large data into meaningful clean data(Kotsiantis & Kanellopoulos, 2006). The collected data contains missing values and categorical data; it has to handle missing values and categorical data to prepare the dataset in a suitable format for the models to use it. After the collection and preparation of the dataset, performing preprocessing which is cleaning the data before they fed into the classifiers has a vital role in accurately and correctly classifying the disease. Following dataset preparation, preprocessing steps are conducted to fed the clean dataset to the machine-learning model:

Cleaning Noisy Data: Cleaning outlier, noisy data is an important part of preprocessing to build a classification model with the best result. Outliers are the values in the dataset, which lies away from the range of the rest of the values. Therefore handling outliers is very important in building a better model. In clinical data, outliers may arise from the natural variance of data. Researchers mostly use boxplots as an easy way to detect outliers. Boxplot uses median and interquartile range IQR to identify the outliers, where the median is the middle number of an ordered set of values and the interquartile range is the variance between the first and third quartiles. To be more precise, outliers are the data points that fall above Q3+1.5(IQR), and below Q1-1.5 (IQR), where Q1 is the first quartile, Q3 is the third quartile and IQR=Q3-Q1**.** The outliers in this study are handled by replacing the value using the mean value.

Handling Missing Values: Some results had missing values due to mistakes while collecting data and some unfilled values of patient's history data. In the actual world, a certain element may be missing for a variety of reasons, including incorrect data, failure to load information, or inadequate extraction. One of the most difficult tasks for analysts is dealing with missing values, because making the proper decision about how to deal with them leads to more robust data models. The classification model's accuracy is affected by missing values. There are several ways of handling missing values namely dropping missing values, filling missing values. To handle these missing values the mean, an average of the observed features used, because all missing features in this study are numeric, and mean imputation is better for numerical missing values.

Handling categorical data: In this step, real data is transformed into the required format. The collected data consists of real data, decimal values, and nominal data. Therefore, in this step nominal data is converted into numerical data of the form 0 and 1. For instance, 'Gender' has the nominal value that can be labeled as 0-for female and 1-for male. After preprocessing the data then the resultant CSV file comprises all the integer and float values for different chronic liver disease related features.

Feature scaling: Is a method for standardizing the independent features included in data over a set of values. It is used to handle drastically changing magnitudes or values during data pre-processing. If feature scaling is not performed, a machine-learning algorithm will weight greater values higher and view smaller values as lower values, independent of the unit of measurement. There are several methods for scaling data, but standard scaling, which scales the descriptive feature to 0 mean and 1 standard deviation, was utilized in this study. Standard scaling is done as follows:

$$Z = \frac{X-\mu}{\sigma} \qquad (3.1)$$

Where:

Z=Standard scale

X=Score or value

μ=Mean and

σ=standard deviation

3.4. Machine learning modeling

This study aims to develop chronic liver disease classification using machine-learning techniques based on already categorized or labeled data. Three supervised machine-learning classification algorithms were used in this study for comparison of the classified models based on their performance. The algorithms have been selected based on their popularity in chronic liver disease classification and their performance of classification on previously done research works. In this study, three machine-learning classification algorithms were selected to develop the model.

Naïve Bayes (NB): A Naive Bayes classifier is a straightforward probabilistic classifier based on the Bayes theorem and a strong independence assumption. The supervised learning approach Naïve Bayes classification assumes that the presence or absence of a given feature in a class is unrelated to the presence or absence of any other feature. The Naive Bayes classifier is a basic and effective classification algorithm that aids in the building of quick machine learning models that can make quick classifications. It's a classification algorithm based on the Bayes theorem. For each class, such as the likelihood of data points linked with a given class, membership probabilities are classified. The Nave Bayes classifier is a conditional probability theorem.

Conditional Probability

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} \quad (3.2)$$

Where:

P (A|B): The evidence's probability if the hypothesis is correct.

P (B|A): The likelihood of the hypothesis if the evidence is correct.

P (A): The chance that hypothesis H is correct. The prior probability is the term for this.

P (B): The evidence's probability.

Support Vector Machine (SVM): This algorithm is the highest prominent and convenient supervised machine-learning algorithm that can be used for data classification, and prediction. Support Vector Machine builds a set of hyperplanes to classify all input in high-dimensional data. In SVM, data is classified into two data points in which hyperplane lies between two classes. SVM creates a discrete hyperplane in the signifier space of the training data and compounds are classified based on the side of the hyperplane located (Jakkula, 2011)(Awad & Khanna, 2015). The hyperplane's position and orientation are defined by support vectors, which are data points that are closer to it. Hyperplanes are a type of decision boundary that aids in the separation of data points. `SVMs were originally designed to classify objects into two categories. The Support Vector Machine is a generalized linear classification algorithm. As a result, the support vector machine can help in hyper-plane discovery.

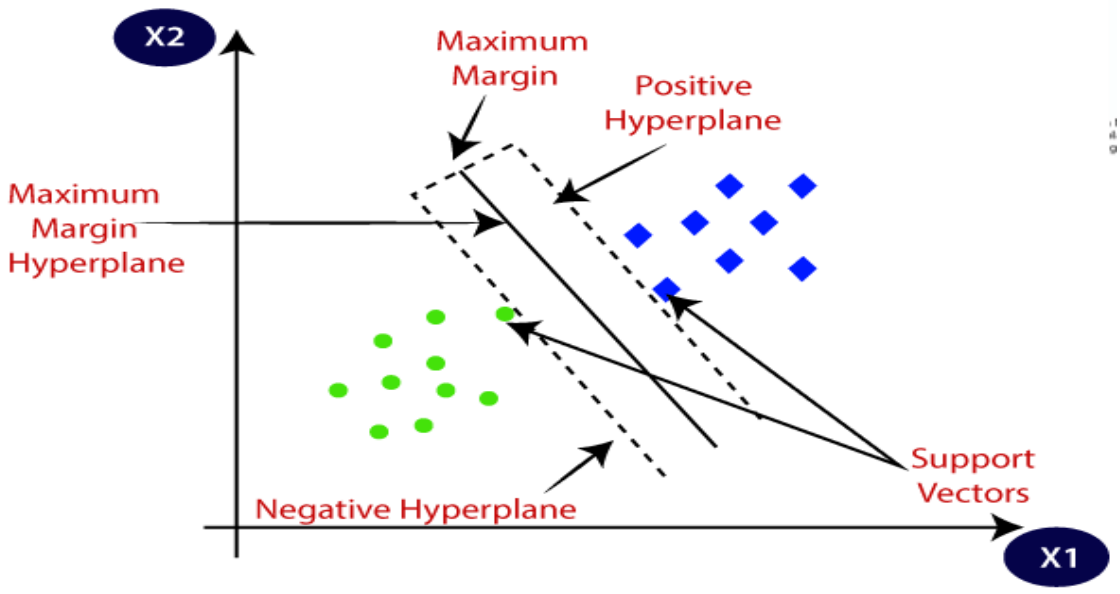

Figure 3. 1 How Support Vector Machine works

The data points that cross through the dotted lines in the preceding above figure are known as support vectors. These support vectors are quite useful in determining whether a data piece is misclassified or not. The term margin refers to the maximum distance between two points. The bigger the marginal distance in linear separable data, the more generalized our model becomes.

Logistic Regression (LR): Logistic regression is a classification algorithm that classifies the output in binary form. One of the most prominent machine learning algorithms used in the Supervised Learning technique is logistic regression. It is a technique for classifying a categorical dependent variable from a collection of independent variables. Logistic regression is used to classify the output of a categorical dependent variable. As a result, the output must be discrete or categorical in nature. It can be Yes or No, 0 or 1, true or false, and so on, but rather than giving exact values like 0 and 1, it gives probabilistic values that are somewhere in between. Logistic Regression is a predictive analytic algorithm based on the concept of probability. It is a Machine Learning technique for categorization problems. Because it can use both continuous and discrete datasets to create probabilities and categorize new data. Logistic regression is a linear classifier, so you will use a linear function.

$$F(x) = b0 + b1x1 + \ldots + brxr \quad (3.2)$$

Where,

F(X) = is logit function,

b0, b1…, br = are samples of estimators of regression coefficients.

r= number of independent features(r=9)

x1, x2…, xr = independent features.

3.4. Evaluation

Performance evaluation is the critical step of developing an accurate machine-learning model. The classification model must be evaluated by the model evaluation techniques to ensure that the model fits the dataset and work well on new unseen input data. The aim of the model performance evaluation is to estimate the generalization accuracy of a model on unseen/out-of-sample data. The performance evaluation method is divided into two categories; holdout and Cross-validation. Holdout evaluation aims to test a model on different data that it was trained on which provides an unbiased estimate of learning performance. The holdout approach for training a machine-learning model entails splitting the data into many splits and utilizing one split for training and the others for validating or testing the models. Both model evaluation and model selection are done using the holdout approach. Holdout is easier to use, more versatile, and faster than cross-validation. However, because variations in the training and test datasets might cause performance variances, this method has a high level of unpredictability(“Mach. Learn. Radiat. Oncol.,” 2015).

Cross-Validation is one of the performance evaluation methods for evaluating and comparing models by dividing data into partitions: one used to train a model and the rest used to test or validate the model. To avoid overfitting, the cross-validation technique is the most used evaluation technique. Cross-validation is a technique that involves dividing the original dataset into a training set, used to train the model, and the test set used to evaluate the model. Therefore, we used the most popular K-fold cross-validation evaluation method to make sure that the models got the pattern for the training dataset for this study. Cross-validation is a technique that involves dividing the original dataset into a training set, used to train the model, and the test set used to evaluate the model. K-fold cross-validation is the basic form of the cross-validation technique. In k-fold validation, the original dataset was partitioned into k equal size subsamples called folds. Therefore, we used the most popular 10-fold cross-validation technique in our study (Belyadi et al., n.d.).

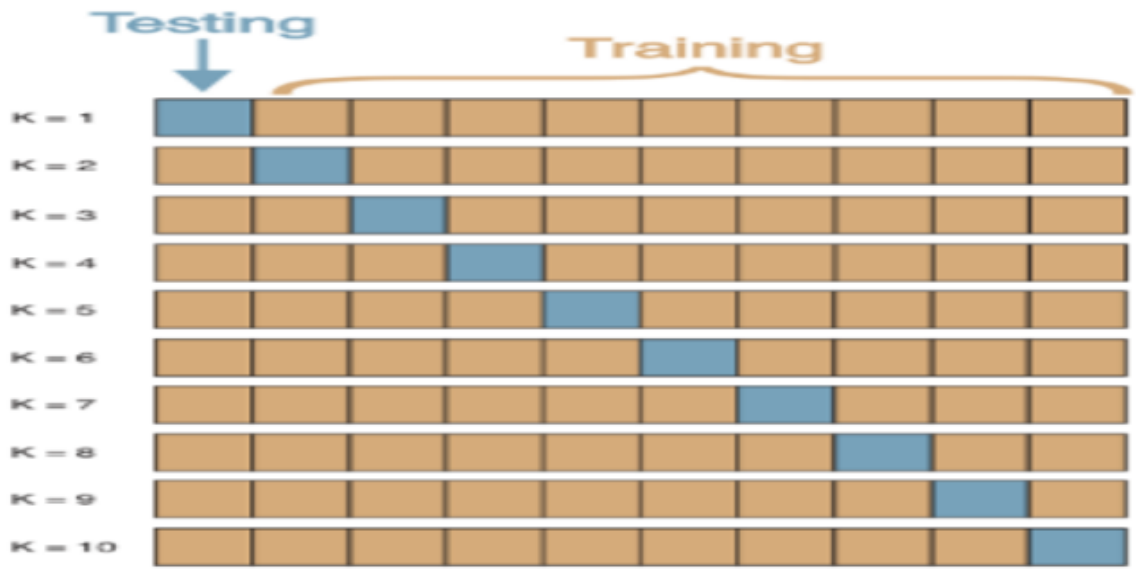

Figure 3. 2: 10-fold cross-validation techniques

3.4.1. Classification model performance evaluation metrics

Classification model evaluation metrics are required to measure model performance because classifier algorithms do not give the same result. Different performance evaluation metrics such as precision, recall, confusion matrix, and f1-score, sensitivity, specificity are used in this study with cross-validation. The followings are the fundamental terms in performance measure.

- True positive (TP): number of examples predicted positive that are actually positive.
- True negative (TN): number of examples predicted negative that are actually negative.
- False positive (FP): number of examples predicted positive that are actually negative.
- False negative (FN): number of examples predicted negative that are actually positive

Accuracy

Accuracy implies the ability of the classification algorithm to classify the classes of dataset correctly. Accuracy is the number of correctly classified data points out of all the data points (Brohi et al., 2019). Generally, accuracy is the measure of the ratio of correct classifies over the total number of instances. Accuracy calculated using the equation:

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}} \quad (3.3)$$

Precision

Precision measure the true values correctly classified from the total classified values in the actual class. Precision quantifies the ability of the classifiers to not label a negative example as positive. The equation to calculate precision is:

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \quad (3.4)$$

Recall

Recall measure the rate of positive values that are correctly classified. Recall answers what proportion of actual Positives is correctly classified.

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \quad (3.5)$$

F-measure

F-measure is also called F1-score is the harmonic mean between recall and precision. The score is calculated by the equation:

$$F1_{score} = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (3.\ 6)$$

Sensitivity

Sensitivity is also called True Positive Rate. Sensitivity is the mean proportion of actual true positives that are correctly identified(Niyati Gupta, 2013). The sensitivity is calculated by the equation:

$$Sensitivity = \frac{TP}{TP + FN} \quad (3.\ 7)$$

Specificity

Specificity is also called True Negative Rate. It is used to measure the fraction of negative values that are correctly classified. The specificity is calculated by the equation:

$$Specificity = \frac{TN}{TN + FP} \quad (3.\ 8)$$

Confusion Matrix

The confusion matrix is a table used to evaluate the performance of classification models for a given set of test data. It can only be determined if the genuine test data values are known. The matrix itself is straightforward, but the terminology used to describe it can be confusing. It's also known as an error matrix since it shows the flaws in the model's performance as a matrix. This study makes use of a confusion matrix and a binary class model (notcld and cld). The binary class confusion matrix is represented in the tables below.

Table 3. 1 Confusion matrix with binary-classes

		Predicted values	
		Cld	Notcld
Actual value	Cld	True cld	False notcld
	Notcld	False cld	True notcld

CHAPTER FOUR

PROPOSED ARCHITECTURE FOR CHRONIC LIVER DISEASE CLASSIFICATION

This chapter discusses the proposed solution for Chronic Liver Disease classification using machine-learning techniques appropriate to solve the identified problem by preparing a chronic Liver disease dataset. Then the machine-learning model is deployed using a flask server. The prototype was developed, to help the practitioners or professionals to diagnose their patients fast and help the patients by giving appropriate advice and treatment based on the status of the patients.

4.1. Proposed Chronic Liver Disease Classification Architecture

Following the identification and validation of the relevant features for the classification of chronic liver disease and the collection of patient history data, the classifier model was built using the machine learning algorithms, namely; Naïve Bayes (NB), Support Vector Machine (SVM), and Logistic Regression (LR). The proposed architecture is used to classify chronic Liver disease as the binary for the classification of cld and notcld. The proposed method takes the dataset as input and then the dataset is preprocessed to make the proposed method complete. Dataset preprocessing process includes data cleaning, handling missing values, handling categorical data. Other evaluation metrics such as precision, recall, f-measure, sensitivity, and specificity are used. The evaluation result is important to select the best model to classify chronic Liver disease. So that, the best model was selected based on the above-mentioned metrics level for the classification of the disease. Finally, the prototype that can take new input data to classify chronic Liver disease is developed using the best-selected classifier model and flask server for deployment. For model development, the most popular and most widely used programming language, which is open source, and free data science libraries are used. Additionally Flask, which is an open-source python package distribution used for web development and Anaconda navigator, consists of different IDEs such as Jupyter, Spyder, PyCharm, and Notebook. The notebook is used for building our model and testing purpose and PyCharm used to work with the web.

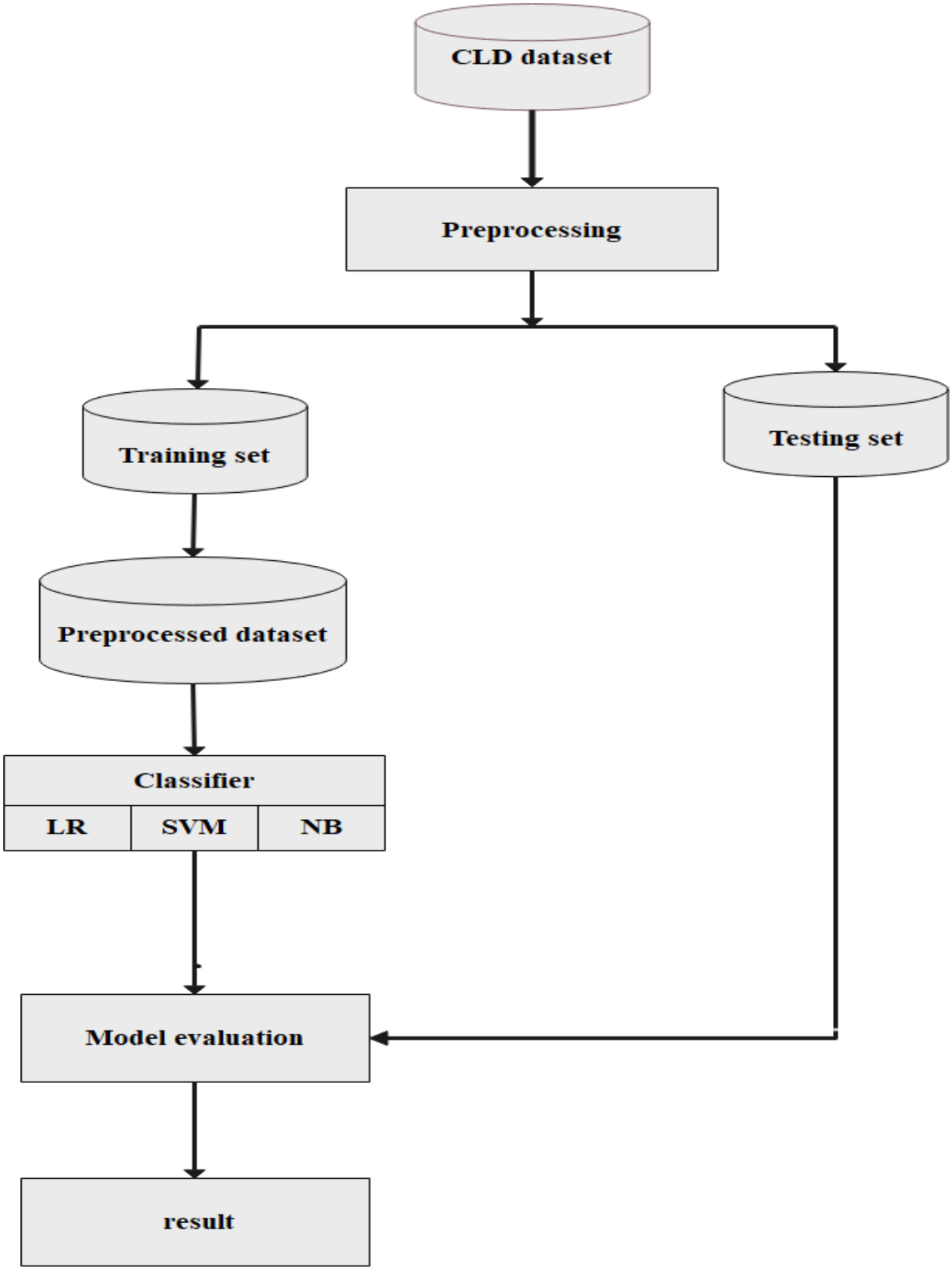

Figure 4. 1 Chronic Liver Disease classification architecture

4.2. Dataset Preparation and Features Description

This study aims to develop a chronic Liver disease classification using machine-learning techniques. Hence, it needs to build a new chronic Liver disease dataset to build the model. This new dataset is needed because currently there is no published or annotated dataset for chronic Liver disease classification in Ethiopia from a local hospital. The process of building the dataset for chronic Liver disease consists of the main steps of gathering data from patient history and prepare the dataset in a suitable way for model building. Preparing the data in a way that is appropriate for the specified machine learning tools, techniques, and tasks is one of the essential machine learning research part. Hence, after the data was collected, the researcher prepared the dataset in a way that the dataset is appropriate to the requirements of the selected machine learning tasks and the specific machine-learning tools. Due to the difficulty of getting a prepared dataset in Ethiopia, data were collected from manually recorded patient history data. Because the data was collected in hardcopy form, it should be converted to softcopy to prepare a machine-readable dataset. As python, programming is used for the development of the model CSV (comma-separated value) file is needed for the code. For the data to be read and design the classification model data should be available in the appropriate format. The chronic liver disease dataset consists of 1619 records in which 832 are records of people with chronic liver disease, and the remaining are records of people without chronic liver disease. The dataset has 9 features based on dataset available in NHR. The last column of the dataset represents the class in which each sample is classified (whether it is a cld or not). A value of 0 indicates that the person has chronic liver disease, and a value of 1 means that the person does not have chronic liver disease. The collected data contains socio-demographic (Age, Gender) features, laboratory test and urine analysis (total bilirubin, direct bilirubin, alkaline_phoshotase, Alamine_aminotransferase, aspartate aminotransferase, albumin and albumin and globulin ratio) features and dependent class (CLD_Status). The prepared dataset had missing values and categorical data that needs to be handled for building classifier models.

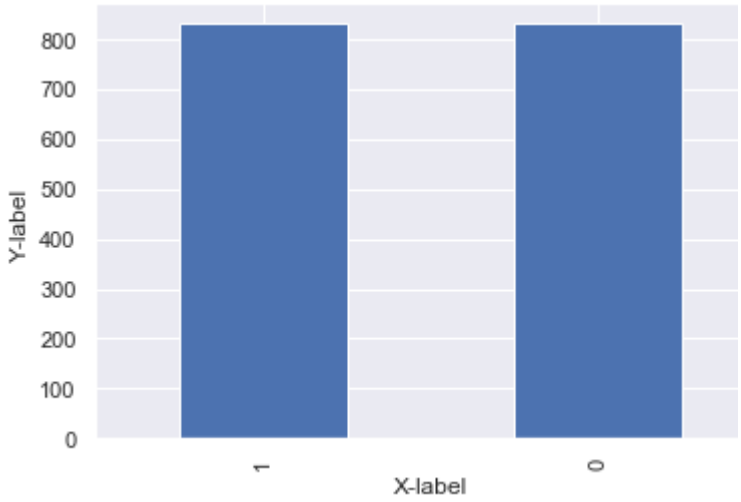

Figure 4. 2 Visualization of chronic liver disease dataset.

Table 4. 1 Dataset Features Description

Symbols	Features full name	Type	Class	Missing values in %
Age	Age	Numeric	Predictor	0
Gender	Gender	Nominal	Predictor	0
TB	Total Bilirubin	Numeric	Predictor	0.11
DB	Direct Bilirubin	Numeric	Predictor	0.49
ALP	Alkaline_phoshotase	Numeric	Predictor	0.43
ALT	Alamine_aminotransferase	Numeric	Predictor	4.07
AST	Aspartate aminotransferase	Numeric	Predictor	0.25
Albumin	Albumin	Numeric	Predictor	0.68
A/G ratio	Albumin and Globulin ratio	Numeric	Predictor	0.802

4.3. Preprocessing Steps

Dataset preprocessing task carried out to prepare proper dataset to be fed into machine learning models to have accurate and correct results. This preprocessing method is performed based on the nature of the dataset. Technique such as cleaning noisy data, filling missing values, are applied to the dataset in this research as shown in figure below:

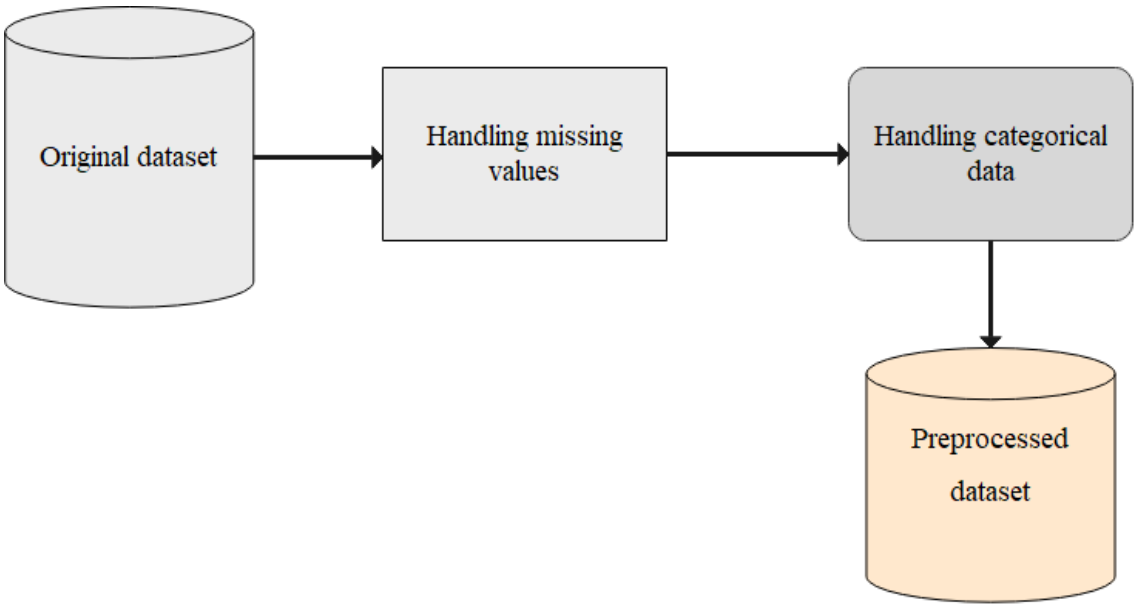

Figure 4. 3 Chronic Liver disease dataset preprocessing steps

4.4 Machine learning Models Building

This study aims to develop a model that enables to classify whether the patient has chronic liver disease or not. This study builds a classification model mainly by using machine-learning models, LR, SVM, and NB. For creating a classifying model, a total an instant of 1619 records was collected for training and testing using learning algorithms. In this study, the classification process outcome is classified from a given input or features relation to class.

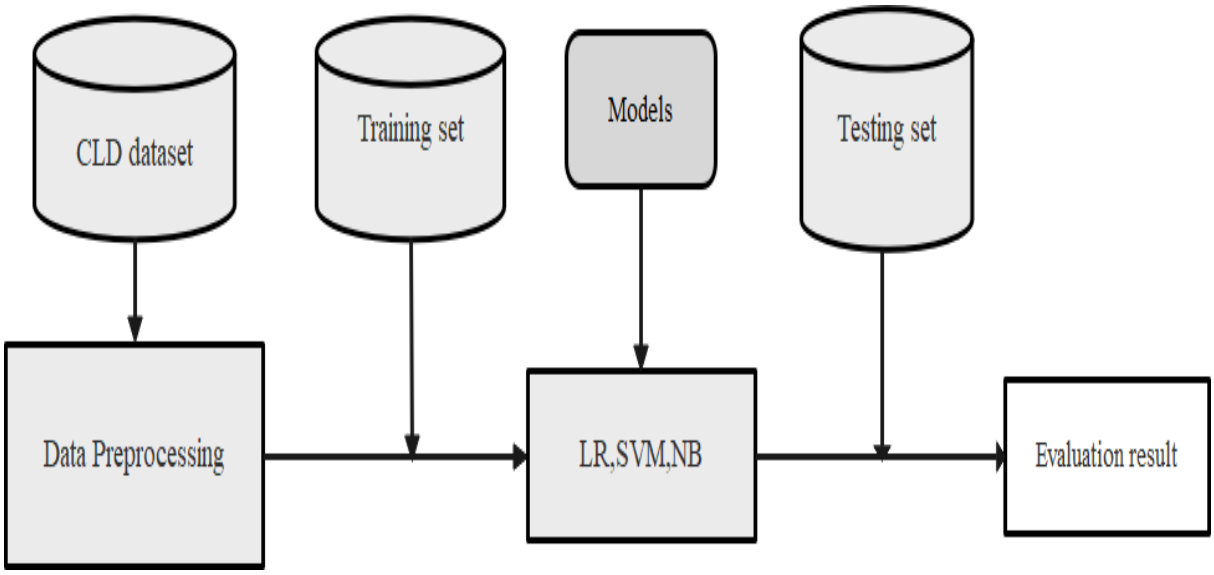

Figure 4. 4 Model Building Flow Diagram

Logistic Regression (LR) is the appropriate methodology to employ when the dependent variable is binary. Logistic regression is a classifier analysis, as are other regression analyses. Logistic regression is used to describe data and explain the relationship between one dependent binary variable and one or more levels of independent variables. A regression model is used to forecast the likelihood that a given input corresponds to the numerical category "1." The sigmoid function is employed in logistic regression to model data. This study also proposed the One-vs-Rest (OVR) strategy of Support Vector Machine (SVM) learning classifier because a simple form of SVM algorithm is a binary classifier. This method separates each class from the rest of the classes in the dataset.

Besides, because LinearSVC is used in this study, one vs rest is an appropriate method with LinearSVC. The third suggested classifier is the Nave Bayes (NB) classifier, which is one of the most simple and effective classification algorithms for constructing fast machine learning models that can make quick classifications. The Naive Bayes technique is a supervised learning algorithm that uses the Bayes theorem to solve classification problems.

4.5 Models Evaluation and Testing

Model evaluation and testing are critical in our study in order to test and estimate the performance of machine learning models trained on the chronic liver disease dataset of binary classes. Since more than two algorithms were used in this study, model evaluation is very important to compare and select the best model for further classification. In this study, the effectiveness of the classifier model is evaluated by using different performance evaluation metrics appropriate for models; these metrics are precision, recall, f1_score, sensitivity, specificity, and confusion matrix are used as discussed in chapter three. The best model is with the best f1_score and accuracy is saved using pickle for deployment on the server after the evaluation.

4.6 Integrating machine-learning model with Flask server.

This study aims to classify chronic liver disease using three models and recommend the best model to be used for deployment on the server after it is evaluated with the best accuracy and f1-score. This is done by preparing the HTML form to insert the new input that is connected to the server and selected model that will allow the professionals to diagnose and give the best treatment and advice for the patient fast and correctly.

CHAPTER FIVE

IMPLEMENTATION AND EXPERIMENTATION

The section describes the implementation of the proposed solution for the chronic Liver disease classification by using the methodology discussed in chapter three and implementing the proposed architecture presented in chapter four. In this study, experiments were done using a binary class dataset. Finally, a prototype for chronic Liver disease classification was developed by using the best model.

5.1. Implementation Environment

In this study, several tools and packages are used to implement the proposed solution. Python programming language has used this study for implementing and experimenting with each proposed solution from the data preprocessing to the final phase, which is the model-building phase and prototype development. The reason to choose python is that it is popular and it is the language of choice for developers, researchers, and data scientists that make the machine learning model development simple and clear. Besides, Python is the general-purpose programming language that can be used for processing text, numbers, images, scientific data easily. It is also simple and easy to understand.

Table 5. 1 Description of the Tools and Python Packages Used During the Implementation

Tools and packages	Version	Description
Anaconda navigator	1.10.0	Help us to launch development applications and easily manage anaconda packages, environments, and channels without the need to use command-line.
Jupyter notebook	6.0.3	It is a free web application that enables us to create and share documents. It is useful for data cleaning and transformation, modeling, data visualization, and machine learning.
Python	3.7.4	Python is a popular platform easy to learn, powerful programming language to develop a machine learning application.
Microsoft Excel	2016	For dataset preparation.
Microsoft Word	2016	For documentation purpose.
Scikit-learn	0.23.1	Is a python module used in machine learning for handling basic ML tasks clustering, regression, and classification?
Pandas	1.0.5	Pandas is an important package providing fast, flexible and expressive data structures designed to make working with “relational” or “labeled” data. It enables integrating and filtering of data, as well as loading it from other external sources like Excel and handling the dataset.
Numpy	1.19.5	Array processing for numbers, strings, and objects. This study uses it for handling converting the text to numeric data for features, training, and testing the model
Matplotlib	3.2.2	Publication quality figures in python. This study uses it for data and results visualization.
Seaborn	0.10.0	Statistical data visualization
Flask server	1.1.1	Microframework for making web services in python. This study uses it for implementing the prototype for selected models

5.2. Deployment Environment

The tools and packages that are discussed above in section 5.1 have been deployed on a personal computer Processor Intel(R) Core(TM) i5-5200U CPU @ 2.2 GHz, 4 Logical Processor(s), 4 Gigabyte of physical memory, 931.51 Gigabyte hard disk storage capacities and Microsoft Windows 10 Enterprise,64bit operating system.

5.3. Dataset Description

In order to build the dataset for this study, we collected the patient history data from manually recorded patient history. First, the researcher assessed different information about the chronic Liver disease in Ethiopia and reviewed different material about the hospital, which works on chronic Liver disease treatment. Then the patient history data from the year 20016 to 2020 was collected by considering the identified attributes for the dataset preparation purpose. Due to a lack of full information and time, not all-patient history data collected. This process results in a total number of 1619 patient data with cld stages and notcld. The dataset had missing values. By applying the data preprocessing method to fill the missing value, the final binary class labeled dataset contains 1619 of data, which labeled with a class of notcld, and cld. The detailed method for building the dataset was discussed in chapter three and chapter four. In addition, the result labeling and size of each class in the dataset are presented in chapter six. Dataset labeling resulted in a distribution of binary class dataset are shown in chapter six. The dataset contains the patient history data in Age, Gender, TB, DB, ALP, ALT, AST, Albumin, A/G ratio along with one of the binary class. The dataset contains 10 columns with the last column as the target class is shown in Appendix A.

5.4. Preprocessing Implementation

Python programming language is used for this study to implement chronic Liver disease dataset preprocessing. The implementation of preprocessing activities includes handling missing values, handling categorical. All none numerical data should be converted to the numerical value. To fill the missing values mean strategy is used and to transform the nominal and none numeric values the study used the Label Encoder and replace method. First, when the dataset is converted into a data frame using pandas, it replaces all blank values with NaN.

```
# importing necessary library and unpreprocessed dataset
 import pandas as pd
 CLD_dataset = pd.read_csv("New chronic liver disease.csv")
```

Figure 5. 1 Python code for loading the panda's library and dataset

5.4.1 Implementation of handling missing values in the dataset

In order to fill the missing values in the dataset, we have used the fillna() method that replaces the missing values by calculating the mean value.

```
CLD_dataset.fillna (CLD_dataset. Mean (), inplace=True)
```

Figure 5. 2 Fill missing value

5.4.2 Implementation of handling categorical data in the dataset

To build a correct and appropriate dataset and to have an accurate model none numeric data should be transformed to numerical data. This data transformation method was implemented using python with the get_dummies module. The method accepts the feature value to be transformed and replace the categorical value with a numerical value.

```
X_encoded = pd.get_dummies(x, prefix_sep = "-"
Y_encoded = Labelencoder().fit_transform(y)
X_scaled = StandardScaler.fit_transform(X_encoded)
```

Figure 5. 3 Handling Categorical data.

5.6 Machine Learning Models Implementation

To build machine-learning models using the proposed algorithm, we used a python sci-kit learning library package and followed the conventional method that is importing the important library package for modeling and metrics for model evaluation.

```
# Importing libraries and packages used for modeling and evaluating the model
    import numpy as np
    import pandas as pd
    from pandas import read_csv
    import matplotlib.pyplot as plt
    from sklearn.metrics import accuracy_score
    from sklearn.model_selection import train_test_split
    from sklearn.metrics import classification_report,confusion_matrix
    from sklearn import linear_model
    from sklearn.linear_model import LogisticRegression
    from sklearn.svm import SVC, LinearSVC
    from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifier,
    BaggingClassifier
    from sklearn.neighbors import KNeighborsClassifier
    from sklearn.naive_bayes import GaussianNB
    from sklearn.linear_model import Perceptron
    from sklearn.tree import DecisionTreeClassifier
    from sklearn.neural_network import MLPClassifier
    from sklearn.preprocessing import StandardScaler
    from sklearn.preprocessing import MinMaxScaler
    from sklearn.preprocessing import RobustScaler
    import matplotlib.pyplot as plt
    import seaborn as sns
```

Figure 5. 4 Importing necessary packages for modeling

After cleaning the dataset, the features split into independent and dependent features. Then the data used as x training and the labels, which contain the class or labels dataset belong set as y. Then, the classifiers train by using the x and y of the entire dataset. To train using the *fit()* method with the appropriate set parameter for each classifier instantiate the class.

```
# Dataset loading and split into independent and depedent
 Cld_dataset= = pd.read_csv("New chronic liver disease.csv")
X = Cld_dataset.drop(columns=”Cld_status”, axis=0)
Y = Cld_dataset[“Cld_status”]
```

Figure 5. 5 Splitting dataset into dependent and independent

The likelihood of a categorical dependent variable is classified using logistic regression. In logistic regression, the dependent variable is a binary variable with data coded as 1 or 0. The logistic regression model predicts P(Y=1) as a function of X. The linear regression model is implemented using the LogisticRegression () technique.

```
# instantiate Logistic Regression model
loreg = LogisticRegression()
loreg.fit(x_train ,  y_train)
```

Figure 5. 6 Instantiate and fit Logistic Regression model

Another proposed classifier in this study is Support Vector Machine. The study used the LinearSVC () classifier of the sklearn package to building the SVM model.

```
# instantiate of Support Vector Machine
linear_svc = SVC(c=100, gamma =0.1)
linear_svc.fit(x_train , y_train)
```

Figure 5. 7 Instantiate and fit SVM model

Another classifier used in the study is the Naïve Bayes classifier. To implement the NB classifier, uses GaussianNB () sklearn package.

```
#instantiate Naïve Bayes Model
NBclassifier = GaussianNB()
NBclassifier.fit(x_train , y_train)
```

Figure 5. 8 Instantiate and fit NB model

5.7 Model Testing and Evaluation

One of the important aspects of machine learning is evaluating the performance of learning algorithms. The evaluation technique implement using the sklearn cross_val_score () and cross_val_predict () methods. These methods use the models and dataset to validate the learning skill of each model.

```
from sklearn.moddel_selection import cross_val_score
Cvscore=cross_val_score(models, x, y, cv=10)
Cvscore
Cvscore.mean()
y_pred = cross_val_predict(models, x, y, cv=10)
```

Figure 5. 9 Applying CV on Models

Additionally in this study, we used classification_report (), confusion_matrix () methods, which are used to evaluate the performance of the built models using classify the value for the entire dataset. These methods give a result of performance evaluation metrics such as precision, recall, and f1-score and other performance metrics like sensitivity and specificity value of the model. Finally, a prototype for classifying chronic Liver disease is implemented using the selected best-performed model then deployed using flask web services for it to be able to accept a new input from the user and then return the result.

CHAPTER SIX

RESULTS AND DISCUSSIONS

This chapter presents the result of the experiment of the proposed chronic Liver disease classfication using machine-learning techniques. In this section, the class distribution of the dataset and experiments to builds classification models for detection based on binary class datasets are discussed. Then we finalize by discussing the result of each experiment in this study and illustrating the graphical user interface of the prototype.

6.1 Data collection result and class distribution results

The total size of the data set after preprocessing is 1619. Out of the total dataset, 832 patients have chronic liver disease (cld), and 787 patients have no chronic Liver disease or normal (notcld). The class distribution in the binary-class dataset is shown in the figures below

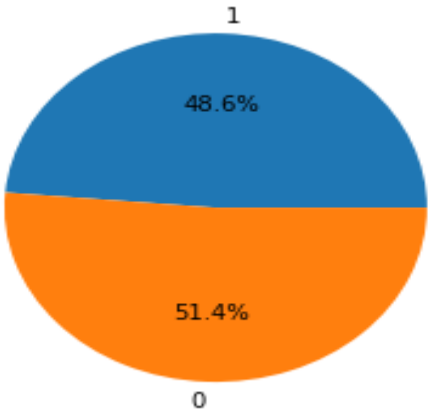

Figure 6. 1 Percentage of class distribution for binary class

Since the binary-class dataset is imbalanced, the data resampling technique was used to balance the label or class of the dataset. The oversampling data resampling technique was used to balance the value of the minority class with the value of the majority class. After the resampling method was used, the size of class labels for 'CLD_Status' class attributes was balanced to 832 cld and 832 notcld and total of 1664 as shown in figure 6.2.

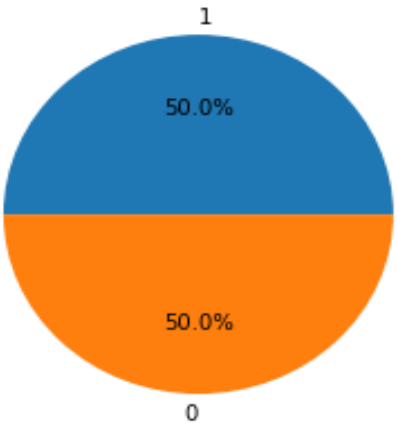

Figure 6. 2 Percentage of class distribution of binary-class dataset after balancing.

Figure 6.2 represents that the class distribution after applying the resampling technique is balanced. The total size of a balanced dataset is 1664.

6.2. Feature Importance

The process of ranking features based on how efficient they are at classifying the target variable is known as feature importance. The feature importance provides a score that indicates how useful each feature in the model. The more important feature in the model will have a higher score of importance. The rank of each feature is based on the comparison of other feature's importance scores in the dataset.

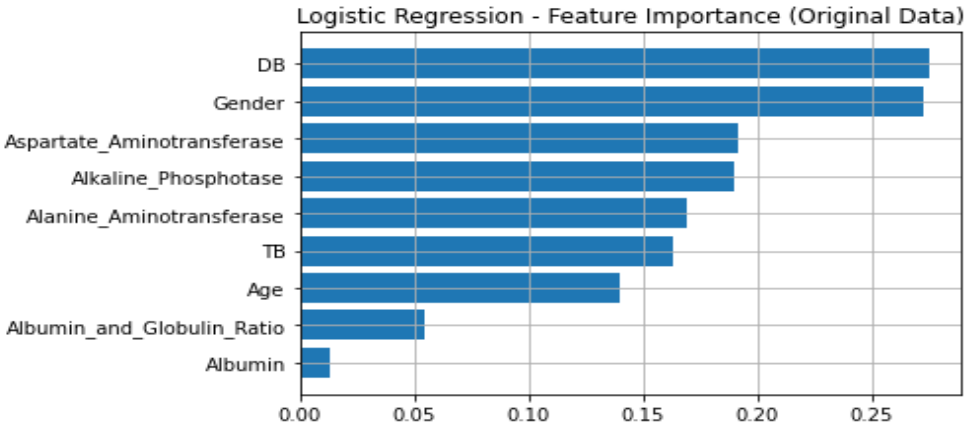

Figure 6. 3 Feature importance using Logistic Regression

6.3. Model Building

In this study, three models namely; Logistic Regression, Support Vector Machine, and Naïve Bayes are proposed. First, the data set is split into independent and dependent. Independent features are the features whose values do not determine by others and the dependent feature is the target or the class whose value depends on the independent features. Then, the models were built on the binary class datasets.

6.4. Model Evaluation Results

The Experiment result is based on three classifiers for binary-class classification. Testing and training these models has performed performance evaluation metrics as discussed in chapter three and chapter four. The obtained results are presented in binary-class classification models. In our thesis work, we conduct the models on the original dataset for binary-class.

6.4.1 Binary Classification Models Evaluation Results on Original Dataset

These classification models built using the two-class dataset that is notcld and cld. The models are trained and tested using performance evaluation metrics. In our experiment, first, we train and test the model on the original dataset. The result of each test accuracy score for LR, SVM and NB models are presented in the tables below as follows.

Table 6. 1 Models accuracy of the binary-class dataset

Models	**Accuracy (%) for LR, SVM, NB**
LR	84
SVM	92
NB	83

Table 6.1 indicates the accuracy score for three classifiers of the binary class dataset. The lowest average accuracy of 83% results was recorded on the NB model, and the highest average accuracy of 92 % resulted in the SVM model. In addition to the accuracy of three classifiers, this study uses other performance evaluation metrics such as precision, recall, f1_score, sensitivity, specificity, and confusion matrix as discussed in chapter three and chapter four. The normalized confusion matrix for each model is described in the following figures to show how correctly the models classify each class along with classifier error.

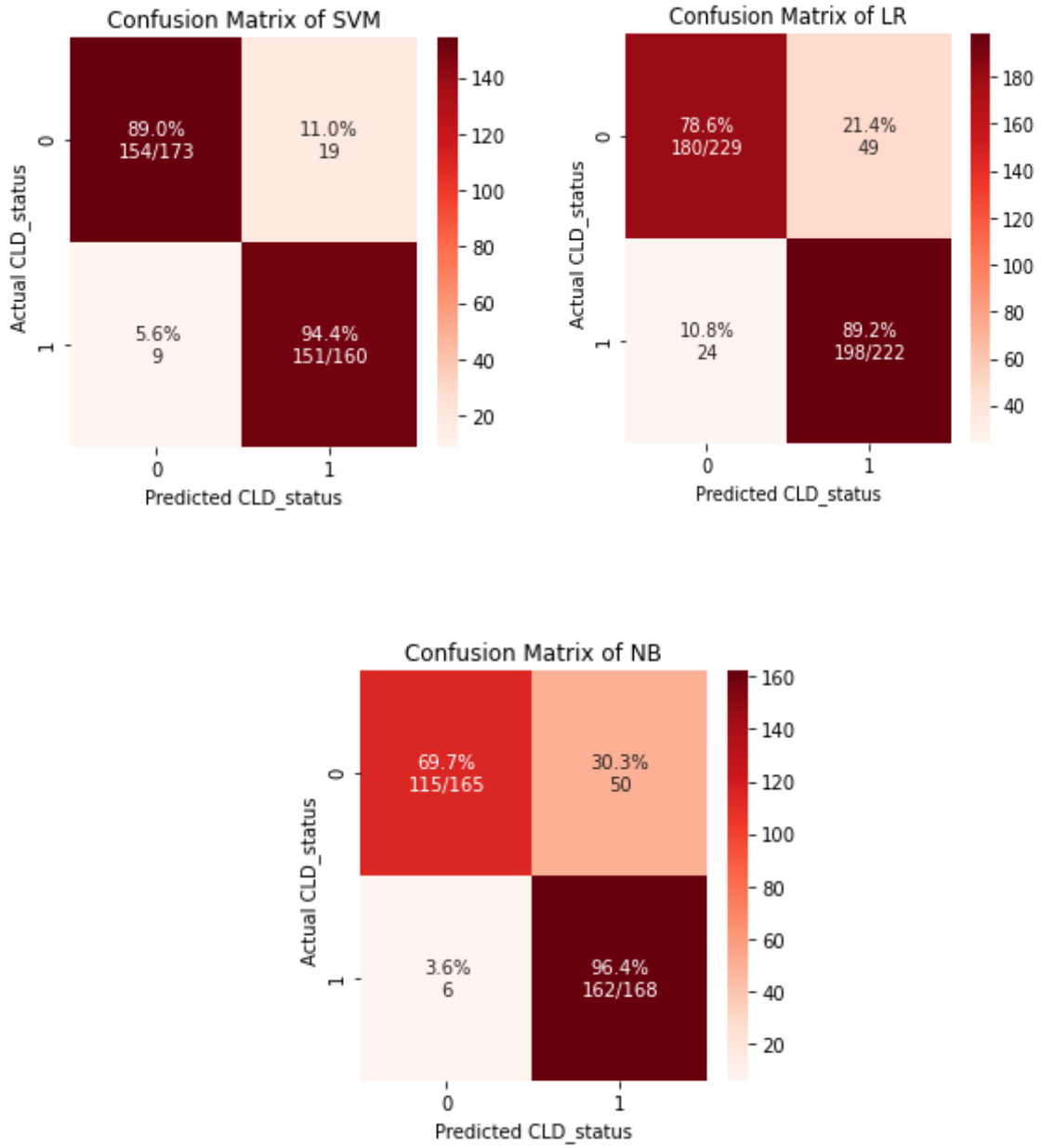

Figure 6. 4 Confusion Matrix of Binary LR, NB, and SVM

Figure 6.4 represents the normalized binary-class confusion matrix for SVM, LR, and NB. SVM classifies 89% of cld, 94.4% of notcld correctly and 11% of cld as notcld and 5.6% of notcld as cld are misclassified. LR classifies 78.6% of cld, 89.2% of notcld correctly and 21.4% of notcld as cld, and 10.8% of notcld as cld are misclassified. NB classify 69.7% of cld, 96.4% of notcld correctly, 30.3% of cld as notcld and 3.6% notcld as cld misclassified. Finally, the result of binary models demonstrates that the SVM model shows better accuracy, than the LR and NB.

Table 6. 2 Performance Result of Binary Classification Model

Models	Performance evaluation metrics				
	Precision	Recall	F1_score	Sensitivity	Specificity
SVM	92	92	92	94.5	89
LR	84	84	84	88	89
NB	83	85	83	83	76

Table 6.2 expresses the performance evaluation metrics results of the three classifiers. Based on the accuracy value, because the dataset is balanced, SVM has a higher score for each performance metric than the rest two models (LR and NB).

6. 5 .Discussions

Chronic liver disease is the most frequent type of liver disease, and middle-aged males were the most affected. Chronic liver disease has become a global challenge, and it has become a silent killer in Ethiopia [49]. Due to a lack of awareness and early diagnosis, many people die or progress to the advanced stages of this disease. As a result, early detection of chronic liver disease can assist to reduce the disease's progression. Early disease detection and classification rely heavily on machine learning. Machine learning serves as a decision support system for medical experts, allowing them to diagnose diseases quickly and effectively.

This study proposed chronic Liver disease classification using machine-learning techniques with a newly prepared dataset. The prepared datasets have 9 features with numerical and nominal values along with the class to which each instance belongs. The prepared dataset had missing values and these missing values were handled using the mean strategy. After preprocessing, we have prepared a binary class. The total dataset size is 1619. The binary class dataset is created by dataset as notcld and cld. After making the binary class, it was an imbalanced dataset and we applied the data resampling technique to balance the dataset.

Three machine-learning models were used in this study. The model evaluation task is done using 10-fold cross-validation and other performance evaluation metrics such as precision, recall, f1_score, sensitivity, and specificity. The confusion matrix is used to show the correctly and incorrectly classified classes to evaluate the performance of the model. The machine learning models used in this study are Logistic Regression, Support Vector Machine, and Naïve Bayes. Several related studies have been done using machine learning and another algorithm in order to classify chronic liver disease globally. Most papers conducted on chronic Liver disease focus on identifying only for the absence or presence of the disease, but do not consider the classification of the severity of the disease.

Shapla Rani Ghosh and Sajjad Waheed(Ghosh & Waheed, 2017) proposed Analysis of classification algorithms for liver disease diagnosis models using Naïve Bayes, Logistic Regression, REP tree, Bagging, and K-star to evaluate the accuracy, precision, and specificity. From these classifications, K-star had maximum accuracy. Finally, K-star algorithms can be used to identify liver disease.

Bilal Khan, Rashid Naseem, Mumtaz Ali, Muhammad Arshad, Nazir Jan (Bilal Khan et al., 2019) proposed classification models, for liver diagnoses based on Composite Hypercube on Iterated Random Projection (CHIRP) by using comparison of different models, such as SVM, KNN,, RF and LR.

Nazim Razali, AidaMustapha, MohdHelmy AbdWahab, Salama A Mostafa, Siti Khadijah Rostam(Razali et al., 2020) proposed classification models of liver disease to produce better performance accuracy by comparing ANN and Naïve Bayes.

G. Ignisha Rajathi and G. Wiselin Jiji (Ignisha Rajathi & Wiselin Jiji, 2019) proposed classification models for chronic liver disease in hybrid Assemble classifier with SVM, KNN and RF. The proposed method employs seven sets of features with a total of 73-3D(three-dimensional) texture features.

Vinutha M.R. Chandrika J(Vinutha & Chandrika, 2021) proposed the method to classify liver disease using Regression Tree, analyzed characteristics of 14 features. The researchers have used the dataset with 435 instances. The dataset is divided into a training dataset of 348 records and remaining 87 records into test data, which is 80% training data and 20% testing data.

Manjunath Varchagall1,D Sivakumar, C Suresh Kumar, Ranjitha V4, Priyanka N (Varchagall et al., 2021) develop a model or classify Chronic Liver Disease in its early stage from datasets containing Liver Function Test (LFT) imbalance results. To balance the datasets, researchers used the Random Forest and K-Nearest Neighbour's (KNN) algorithms.

The proposed model of this study was evaluated using 10-fold cross-validation and other performance evaluation metrics such as precision, recall, f1_score, sensitivity, and specificity. In this study, different methodologies have been conducted to do an experiment and analysis the result. Three machine-learning models, LR, SVM, and NB have been implemented, and SVM records accuracy of 92%, LR records accuracy of 84%, and NB records accuracy of 83%. The model with high accuracy to identify the disease fast and accurately and reduce the economic burden of patients. Thus, SVM is recommended in our study based on its accuracy, f1_score, and other performance evaluation on the original dataset. Table 6.3 shows the comparison of the proposed work and previously done works on chronic liver disease classification.

Table 6. 3 Comparison of proposed model and previous works

Proposed Model	**Methodology**	**Accuracy result (%)**
	SVM	**92**
Related works	Methodology	Accuracy result (%)
[Ghosh & Waheed, 2017] Analysis of classification algorithms for liver disease diagnosis	NB, and LR	LR with 75.4%
[(Bilal Khan et al., 2019] Machine learning approaches for diagnosis of liver disease	SVM, KNN, , RFnd LR	RF with 72.12%
[Razali et al., 2020] Data Mining Approaches to the prediction of liver disease	ANN and NB	NB with 70.5%
[(Ignisha Rajathi & Wiselin Jiji, 2019] Chronic Liver Disease Classifications Assemble classifier	SVM, KNN and RF	SVM with 84%
[Varchagall et al., 2021] Early Forecasting Chronic Liver Disease from Liver Function Test Imbalance Datasets	RF and KNN	KNN with 75.96

This study used an original dataset with 9 features and one target class as clinical classification to classify the severity of the disease and developed a web app prototype. The disease severity can be quickly classified to assist the physician in making decisions and providing patients with further appropriate examination and treatment.

6.5.1 User Interface

The user interface is the means of communication that enables the end-user (experts) to easily access the system with a prepared prototype using HTML. The server either initialized from the command prompt or using PyCharm IDE.

After initializing the server it gives the HTTP address which contains the Html page from which the end-user insert the value of the feature and see the result as shown in fig 6.5.

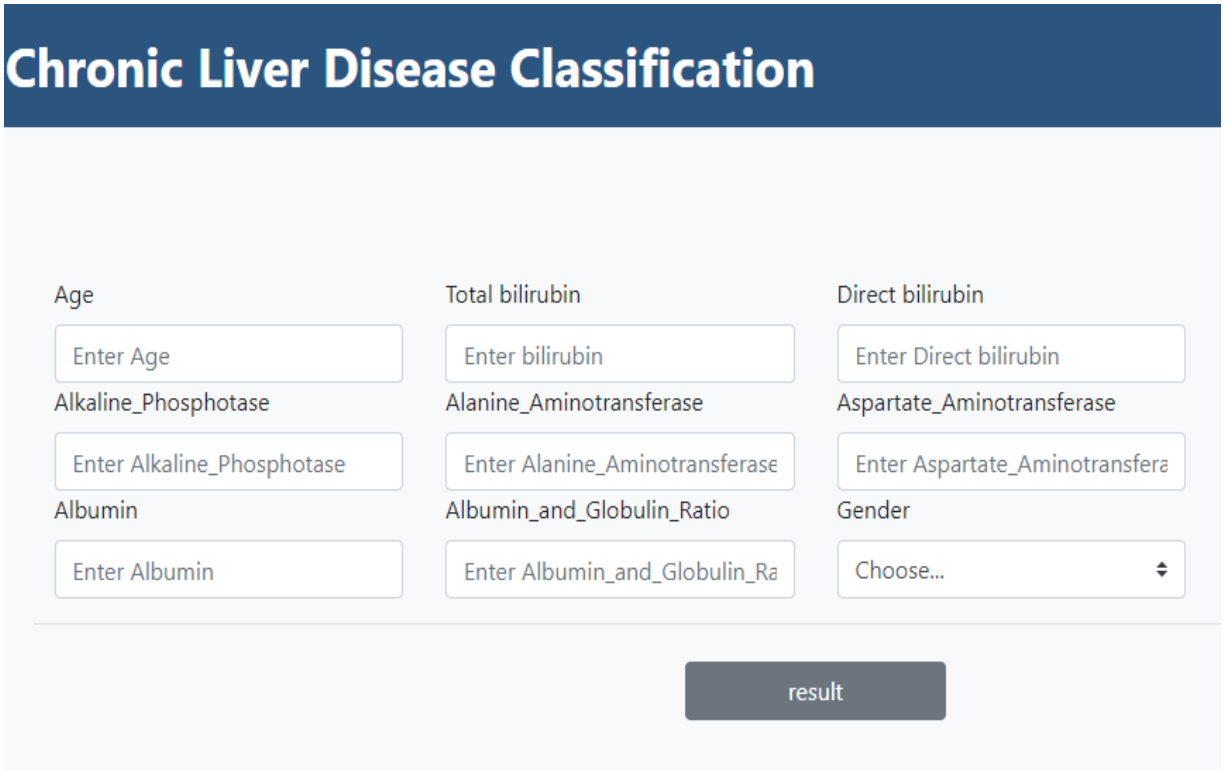

Figure 6. 5 GUI for data input

CHAPTER SEVEN

CONCLUSION RECOMMENDATION AND FUTURE WORK

In this section, the classification system for chronic liver disease is proposed with the aid of machine learning. The technical results, recommendations, and future orientations of the research are outlined.

7.1 Conclusion

Classification is the primary data retrieval technique that is mainly used in the health sector for medical diagnosis and disease classification. In this study, the chronic liver disease classification system using the machine-learning technique is proposed with the deployment of the model to help the experts to diagnose the disease fast. Pre-processing measurements were taken in this study to make the data clean and adapted to the machine-learning model using machine learning and the data science tool that is python. In this study, the missing values have been handled using the mean replacement method and categorical values have been converted to binary, which was assigned number using the replace method. The resample technique was used to balance the dataset for binary class. The cleaned total dataset size is 1619 and the balanced and cleaned dataset for the binary class is 1664. Three-machine learning models LR, SVM, NB were used to build the proposed model. The model evaluation task is done using 10-fold cross-validation. Applying the models on the original dataset, we have the highest accuracy with an SVM scored accuracy of 92%.

7.2. Recommendation

These days, chronic liver disease has become a challenge around the world and many people are dying of this disease in Ethiopia even without knowing they had a chronic liver disease previously. Thus, the application of machine learning techniques in clinical data of chronic Liver disease is an important research area to improve the services provided as well as to come up with solutions to prevent and slow down the progress of this disease. In Ethiopia, there is no integration of health care system and technology, which leave a huge burden for the health experts and difficulties for the researchers to do their research, which needs more improvement in the health center for the experts to give appropriate service for the patients and for the researchers to do their work properly and efficiently.

Based on the challenges and finding of this study the researcher recommend the tasks that need to be improved for future works especially for machine learning and deep learning related medical research work. Finding organized data is challenging, data has been transformed from a paper file to an electronic format, and it was a boring and time-consuming task. Thus, the health care organization should try to store patient history data in a computer system that enables the experts and concerned body to retrieve the data easily when it is needed. Though the dataset used for this study purpose is big when compared to the real-world chronic liver disease dataset, it is better to implement the models using a dataset with more instances from various sources. Thus, it is recommended that these models should be applied and proved to this data**.**

7.3. Future Works

This study used a supervised machine-learning algorithm. It is preferable to see the difference in performance outcomes by using models of unsupervised or deep learning algorithms. The proposed model supports the experts to give decisions on, it is better to make it a mobile-based system that enables the experts to follow the status of the patients and help the patients to use the system to know their status. In addition, the model requires further improvement regarding feature selection of the chronic liver disease into multiple components. It is preferable to work with a real-time dataset for greater accuracy.

Special Acknowledgement: This research was funded by Adama Science and Technology University under the grant number **ASTU/SM-R/233/21.**

Adama, Ethiopia

REFERENCES

Aravind, A., Bahirvani, A. G., Quiambao, R., & Gonzalo, T. (2020). Machine Learning Technology for Evaluation of Liver Fibrosis, Inflammation Activity and Steatosis (LIVERFASt^{TM}). *Journal of Intelligent Learning Systems and Applications*, *12*(02), 31–49. https://doi.org/10.4236/jilsa.2020.122003

Awad, M., & Khanna, R. (2015). Efficient learning machines: Theories, concepts, and applications for engineers and system designers. *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*, *July 2018*, 1–248. https://doi.org/10.1007/978-1-4302-5990-9

Banait, S., Badole, S. M., Jain, J., & Thorat, A. (2021). Risk factors for chronic liver disease in population of Central India: a case-control study from rural India. *Egyptian Liver Journal*, *11*(1). https://doi.org/10.1186/s43066-021-00077-9

Bansal, H., Sharma, K., & Kaur, G. (2020). *A HYBRID APPROACH FOR THE PREDICTION OF LIVER DISORDER USING DATA MINING*. *9*(7), 164–171.

Banu Priya, M., Laura Juliet, P., & Tamilselvi, P. R. (2018). Performance Analysis of Liver Disease Prediction Using Machine Learning Algorithms. *International Research Journal of Engineering and Technology(IRJET)*, *5*(1), 206–211.

Belyadi, H., Haghighat, A., Using, G., Threat, I., & Systems, R. (n.d.). *Evaluation Metric*.

Blundell, R., Azzopardi, J. I., Liver, N. F., Disease, L., Injury, L., & Cirrhosis, L. (n.d.). *Chronic Liver Disease*.

Bollyky, T. J., Templin, T., Cohen, M., & Dieleman, J. L. (2017). Lower-income countries that face the most rapid shift in noncommunicable disease burden are

also the least prepared. *Health Affairs*, *36*(11), 1866–1875. https://doi.org/10.1377/hlthaff.2017.0708

Brohi, S. N., Pillai, T. R., Kaur, S., Kaur, H., Sukumaran, S., & Asirvatham, D. (2019). Accuracy Comparison of Machine Learning Algorithms for Predictive Analytics in Higher Education. *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, *285*(July), 254–261. https://doi.org/10.1007/978-3-030-23943-5_19

Byass, P. (2014). *The global burden of liver disease : a challenge for methods and for public health*. 10–13.

Cheemerla, S., & Balakrishnan, M. (2021). Global Epidemiology of Chronic Liver Disease. *Clinical Liver Disease*, *17*(5), 365–370. https://doi.org/10.1002/cld.1061

Durai, V. (n.d.). *Liver disease prediction using machine learning*. *5*(2), 1584–1588.

Ghosh, S. R., & Waheed, S. (2017). Analysis of classification algorithms for liver disease diagnosis. *Journal of Science Technology and Environment Informatics*, *5*(1), 361–370. https://doi.org/10.18801/jstei.050117.38

Ignisha Rajathi, G., & Wiselin Jiji, G. (2019). Chronic liver disease classification using hybrid whale optimization with simulated annealing and ensemble classifier. *Symmetry*, *11*(1). https://doi.org/10.3390/sym11010033

Jacob, J., Chakkalakal Mathew, J., Mathew, J., & Issac, E. (2018). Diagnosis of Liver Disease Using Machine Learning Techniques. *International Research Journal of Engineering and Technology*, *5*(4), 4011–4014.

Jakkula, V. (2011). Tutorial on Support Vector Machine (SVM). *School of EECS,*

Washington State University, 1–13.

Jeong, S., Yang, H., & Hwang, H. (2017). Evaluation of an automated connective tissue disease screening assay in Korean patients with systemic rheumatic diseases. *PLoS ONE*, *12*(3), 1–14. https://doi.org/10.1371/journal.pone.0173597

Jin, H., Kim, S., & Kim, J. (2014). Decision factors on effective liver patient data prediction. *International Journal of Bio-Science and Bio-Technology*, *6*(4), 167–178. https://doi.org/10.14257/ijbsbt.2014.6.4.16

Khan, Bilal, Naseem, R., Ali, M., Arshad, M., & Jan, N. (2019). Machine learning approaches for liver disease diagnosing. *International Journal of Data Science and Advanced Analytics (ISSN 2563-4429)*, *1*(1), 27–31.

Khan, Binish, Shukla, P. K., & Ahirwar, M. K. (2019). Strategic Analysis in Prediction of Liver Disease Using Different Classification Algorithms. *International Journal of Computer Sciences and Engineering*, *7*(7), 71–76. https://doi.org/10.26438/ijcse/v7i7.7176

Khan, Binish, Shukla, P. K., Ahirwar, M. K., & Mishra, M. (2020). *Strategic Analysis in Prediction of Liver Disease Using Different Classification Algorithms*. *7*, 437–449. https://doi.org/10.4018/978-1-7998-2742-9.ch022

Kohli, P. S., & Regression, A. L. (2015). International Conference on Computing, Communication and Automation, ICCCA 2015. *International Conference on Computing, Communication and Automation, ICCCA 2015*, 1–4.

Kotsiantis, S. B., & Kanellopoulos, D. (2006). Data preprocessing for supervised leaning. *International Journal of …*, *1*(2), 1–7. https://doi.org/10.1080/02331931003692557

Krishnan, S. (2019). *LIVER DISEASES-AN OVERVIEW*. *March*. https://doi.org/10.20959/wjpps20191-13036

Liu, J., Zhang, Z., & Razavian, N. (2018). Deep EHR: Chronic disease prediction using medical notes. *ArXiv*.

Machine Learning in Radiation Oncology. (2015). In *Machine Learning in Radiation Oncology*. https://doi.org/10.1007/978-3-319-18305-3

Magnus, S., & Orlien, S. (2019). *Catha edulis chronic liver disease in eastern Ethiopia*.

Mehtaj Banu, H. (2019). Liver disease prediction using machine-learning algorithms. *International Journal of Engineering and Advanced Technology*, *8*(6), 2532–2534. https://doi.org/10.35940/ijeat.F8365.088619

Niyati Gupta, N. G. (2013). Accuracy, Sensitivity and Specificity Measurement of Various Classification Techniques on Healthcare Data. *IOSR Journal of Computer Engineering*, *11*(5), 70–73. https://doi.org/10.9790/0661-1157073

Online, F. A., Bagny, A., Bouglouga, O., Balaka, A., Atakouma, Y. D., & Redah, D. (2021). *Journal of Spleen and Liver Research*. *1*, 1–7. https://doi.org/10.14302/issn.2578

Pingale, K., Surwase, S., Kulkarni, V., Sarage, S., & Karve, P. A. (2019). *Disease Prediction using Machine Learning*. 831–833.

Razali, N., Mustapha, A., Wahab, M. H. A., Mostafa, S. A., & Rostam, S. K. (2020). A Data Mining Approach to Prediction of Liver Diseases. *Journal of Physics: Conference Series*, *1529*(3). https://doi.org/10.1088/1742-6596/1529/3/032002

Sivakumar, D., Varchagall, M., Ambika, L. G., & Usha, S. (2019). Chronic liver

disease prediction analysis based on the impact of life quality attributes. *International Journal of Recent Technology and Engineering*, *7*(6), 2111–2116.

Sontakke, S., Lohokare, J., & Dani, R. (2017). Diagnosis of liver diseases using machine learning. *2017 International Conference on Emerging Trends and Innovation in ICT, ICEI 2017*, 129–133. https://doi.org/10.1109/ETIICT.2017.7977023

Technology, M. O. F. (2017). *APPLICATION OF MACHINE LEARNING ALGORITHMS TO A WELL DEFINED CLINICAL PROBLEM : LIVER DISEASE*.

Tesfaye, B. T., Feyissa, T. M., Workneh, A. B., Gudina, E. K., & Yizengaw, M. A. (2021). *Chronic Liver Disease in Ethiopia with a Particular Focus on the Etiological Spectrums : A Systematic Review and Meta-Analysis of Observational Studies*. *2021*.

Thirunavukkarasu, K., Singh, A. S., Irfan, M., & Chowdhury, A. (2018). Prediction of liver disease using classification Algorithms. *2018 4th International Conference on Computing Communication and Automation, ICCCA 2018*, *6*(9), 1–3. https://doi.org/10.1109/CCAA.2018.8777655

Varchagall, M., Sivakumar, D., Sureshkumar, C., Ranjitha, V., & Priyanka, N. (2021). *Early Forecasting of Chronic Liver Disease from Liver Function Test Imbalance Datasets*. *25*(4), 2704–2716.

Vijayarani, D. (2015). Liver Disease Prediction using SVM and Naïve Bayes Algorithms. *International Journal of Science, Engineering and Technology Research*, *4*(4), 816–820.

Vinutha, M. R., & Chandrika, J. (2021). Prediction of Liver Disease using Regression Tree. *International Journal of Online and Biomedical Engineering*, *17*(2), 164–172. https://doi.org/10.3991/ijoe.v17i02.19287

WHO. (2020). *World Health Organization. Package of Essential Noncommunicable (PEN) Disease Intervention for Primary Health Care. WHO*.

APPENDICES

Appendix A: Dataset Features Description

- **Age:** Age of the patient
- **Gender:** is the gender of the patient (Male or Female)
- **TB:** In milligrams per deciliter (mg/dL), total bilirubin is measured. Bilirubin is an orange-yellow pigment produced by the breakdown of a portion of your red blood cells. The amount of bilirubin in your blood is measured by a bilirubin test. It is used to figure out what's causing health problems like jaundice, anemia, and liver illness.
- **DB:** Milligrams per deciliter (mg/dL) of direct bilirubin. Direct or conjugated bilirubin is bilirubin that has been linked to glucuronic acid, a glucose-derived acid, by the liver. Indirect bilirubin is bilirubin that is not connected to glucuronic acid.
- **ALP:** Units per liter (U/L) of alkaline phosphatase. Alkaline phosphatase (ALP) is a blood enzyme that aids in the breakdown of proteins.
- **ALT:** Alanine aminotransferase, often known as SGPT, is measured in units per liter (U/L). The enzyme alanine aminotransferase (ALT) is mostly present in the liver and kidney. ALT is used to screen for and/or monitor liver disease because it increases with liver damage.
- **AST**: Units per liter (U/L) of aspartate aminotransferase, often known as SGOT. The enzyme AST (aspartate aminotransferase) is primarily located in the liver, although it can also be found in the muscles. AST is released into the bloodstream when your liver is damaged.
- **ALB:** Albumin is measured in g/dL (grams per deciliter). The most prevalent protein present in the bloodstream is albumin. Its primary job is to maintain osmotic pressure, which is the process that keeps fluids from seeping out of blood vessels and into the surrounding tissues.
- **AGR:** Ratio of albumin to globulin. A total protein test, which uses a blood sample to evaluate the total combined amount of albumin and globulin in the blood, yields the A/G ratio.
- **Cld_Status:** Status of the Patient (patient with cld [0], or notcld [1])

Appendix B: Sample Code

Appendix B.1 Sample Code for Preprocessing

#Preprocessing

```
from sklearn.preprocessing import LabelEncoder,OneHotEncoder

#Handling Categorical data

X = pd.get_dummies(X, prefix_sep='_')

Y = LabelEncoder().fit_transform(Y)

#Filling missing values

from sklearn.impute import SimpleImputer

X = SimpleImputer(missing_values= np.NAN, strategy= 'mean', fill_value=None,

verbose=0, copy=True)

X =X.fit(X)
```

Appendix B.2 Sample source code for Building and Evaluating Models

LR model

```
# Create logistic regression object

logreg = LogisticRegression()

# Train the model using the training sets and check score

logreg.fit(X_train, y_train)

#Predict Output

log_predicted= logreg.predict(X_test)

logreg_score = round(logreg.score(X_train, y_train) * 100, 1)

logreg_score_test = round(logreg.score(X_test, y_test) * 100, 1)

#Equation coefficient and Intercept
```

```
print('Logistic Regression Training Score: \n', logreg_score)

print('Logistic Regression Test Score: \n', logreg_score_test)

print('Coefficient: \n', logreg.coef_)

print('Intercept: \n', logreg.intercept_)

print('Accuracy: \n', accuracy_score(y_test,log_predicted))

print('Confusion Matrix: \n', confusion_matrix(y_test,log_predicted))

print('Classification Report: \n', classification_report(y_test,log_predicted))

#dataframe = pd.DataFrame(confusion_matrix, index=liver_df, columns=liver_df)

sns.heatmap(confusion_matrix(y_test,log_predicted),annot=True,cbar=None,
fmt="d",color='red')

plt.title("Confusion Matrix"), plt.tight_layout()

plt.ylabel("Actual Cld_status"), plt.xlabel("Predicted  Cld_status")

plt.show()
```

SVM model

```
linear_svc=SVC()

linear_svc.fit(X_train, y_train)

#Predict Output

svm_predicted = linear_svc.predict(X_test)

svm_score = round(linear_svc.score(X_train, y_train) * 100, 2)

svm_score_test = round(linear_svc.score(X_test, y_test) * 100, 2)

print('SVM Score: \n', svm_score)

print('SVM Test Score: \n', svm_score_test)

print('Accuracy: \n', accuracy_score(y_test, svm_predicted))

print(confusion_matrix(y_test,svm_predicted))

print(classification_report(y_test,svm_predicted))
```

```
#plot_confusion_matrix(model, X_test, y_test)

sns.heatmap(confusion_matrix(y_test,svm_predicted
),annot=True,fmt="d",cmap='YlGnBu')
```

 # NB model

```
# Gaussian Naive Bayes

gaussian = GaussianNB()

gaussian.fit(X_train, y_train)

#Predict Output

gauss_predicted = gaussian.predict(X_test)

gauss_score = round(gaussian.score(X_train, y_train) * 100, 2)

gauss_score_test = round(gaussian.score(X_test, y_test) * 100, 2)

print('Gaussian Score: \n', gauss_score)

print('Gaussian Test Score: \n', gauss_score_test)

print('Accuracy: \n', accuracy_score(y_test, gauss_predicted))

print(confusion_matrix(y_test,gauss_predicted))

print(classification_report(y_test,gauss_predicted))

#sns.heatmap(confusion_matrix, annot=True, fmt='d', cmap='YlGnBu')

#sns.heatmap(dataframe, annot=True, cbar=None, cmap="Blues")

sns.heatmap(confusion_matrix(y_test,gauss_predicted),annot=True,fmt="d",cmap='YlGnB
u').
```

Appendix B.3 Sample Code for model saving

 # Saving model SVM

```
linear_svc=SVC(C=100, gamma=0.1)

linear_svc.fit(X_train, y_train)

svm_predicted = linear_svc.predict(X_test)

svm_score = round(linear_svc.score(X_train, y_train) * 100, 2)

svm_score_test = round(linear_svc.score(X_test, y_test) * 100, 2)

print('SVM Score: \n', svm_score)

print('SVM Test Score: \n', svm_score_test)

print('Accuracy: \n', accuracy_score(y_test, svm_predicted))

pickle.dump(SVM model, open('SVM _model.pkl', 'wb'))
```

Appendix B.4 Sample code for integrating ML model with flask server

#chronic liver disease classification web app

```
import numpy as np

import pickle

from flask import Flask, render_template,request

# Create flask app

flask_app = Flask(__name__)

model = pickle.load(open("SVM_model.pkl", "rb"))

#@app = Flask(__name__)

@flask_app.route('/')

def home():
```

```
    return render_template('index.html')
@flask_app.route("/result", methods=['POST','GET'])
def result():
    features= [float(x) for x in request.form.values()]
    final_features = [np.array(features)]
    #features_name = ['anr', 'sd', 'k','texture','mint', 'avp','cec','ph','maxt','tn','crop','om' ]
   # df = pd.DataFrame(feature_value)
    result= model.predict(final_features)
    result=""
    if result == [0]:
        result = "The patient has Chronic Liver Disease!"
    else if result == [1]:
       result = "The patient has no Chronic Liver Disease!"
    return render_template("index.html",Final_result =result)
if __name__ == '__main__'
     flask_app.run()
```