LIVER CIRRHOSIS PREDICTION USING MACHINE LEARNING
BY

Mahabub Alam Shakil
ID: 193-15-13544

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

Supervised By

Md. Zahid Hasan
Associate Professor and
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JULY, 2024

APPROVAL

This Project titled **"Liver Cirrhosis Prediction Using Machine Learning Algorithm"**, submitted by **Mahabub Alam Shakil ID:193-15-13544** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **14 July 2024.**

BOARD OF EXAMINERS

Dr. S.M Aminul Haque **Chairman**
Professor & Associate Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Md. Abbas Ali Khan **Internal Examiner**
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Mr. Md. Aynul Hasan Nahid **Internal Examiner**
Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

14.07.2024

Dr. Abu Sayed Md. Mostafizur Rahaman **External Examiner**
Professor
Department of computer science and engineering
Jahangirnagar University

DECLARATION

I hereby declare that, I have done this project under the supervision of **Md. Zahid Hasan,** Associate Professor, Department of CSE, Daffodil International University. I further declare that neither an application nor an educational grant has been made anywhere for this project or any part of it.

Supervised by:

Md. Zahid Hasan
Associate Professor
Department of CSE
Daffodil International University

Submitted by:

Mahabub

Mahabub Alam Shakil
ID: 193-15-13544
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First I express my heartiest thanks and gratefulness to almighty God for His divine blessing makes it possible to complete the final year project successfully.

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

I would like to express my heartiest gratitude to Professor **Dr. Sheak Rashed Haider Noori,** Head**,** Department of CSE, for his kind help to finish my project and also to other faculty members and the staff of CSE department of Daffodil International University.

I would like to thank all of my classmates at Daffodil International University who participated in the discussion while also attending the course.

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

ABSTRACT

Liver cirrhosis is a highly infectious blood-borne illness that is often asymptomatic in its early stages. Consequently, identifying and curing patients at an early stage of the ailment is difficult. As the disease spreads to its advanced stages, diagnosis and treatment become even more difficult. The goal of this research is to propose an artificial intelligence system utilizing machine learning algorithms conservatively created to support physicians in diagnosing liver cirrhosis at an early stage. To estimate the likelihood of a liver cirrhosis infection, it was decided to construct a model for training produced machine learning algorithms that use a variety of physiological variables. Three models for accurate prognostication have been developed with different training, linking three different sets of physiological variables and machine learning algorithms built on LR, SVM, DTC, GB and RFC. The algorithm that performed the best in this challenge was Random Forest, which had an accuracy of almost 86.74%. The approach was developed using the publicly-accessible Liver Cirrhosis data source. The models used in this study had a significantly higher accuracy than those used in previous studies, indicating their increased dependability. Their resilience has been demonstrated by several model comparisons, and the research study may be used to select the scheme.

Keywords: Liver cirrhosis, machine learning, predictive modeling, artificial intelligence, classification algorithms.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Introduction

The illness known as liver cirrhosis affects every person on the planet. It is a blood-borne disease that is transmitted by direct contact with the blood or body fluids that include blood of infected individuals. This disorder causes chronic illness in around 71 million people globally, and it is projected to have killed 399,000 people in 2016 [1]. The World Health Organization (WHO) states that liver cirrhosis is a global disease. The World Health Organization's study indicates that 3-5 million people contract this virus annually. The highest prevalence of this virus is seen in developing Asian and African countries, as opposed to wealthier European and North American ones. In addition, the number of individuals with chronic diseases is rising in nations like Egypt, China, and Pakistan [2-4]. The signs of liver cirrhosis virus manifest significantly later in the course of the illness. When an infection is caught in its early stages, almost 80% of affected individuals do not show any symptoms, which increases liver damage and death rates.

As far as vaccinations against the liver cirrhosis virus go, none are effective. Therefore, assessing the degree of liver damage sustained by the affected patient may help medical professionals diagnose and effectively treat chronic infections. Since proper management stops viruses from spreading between people, it is essential for preventing illness [5-7]. Developments in artificial intelligence (AI) help physicians diagnose and treat patients more quickly. Research has compared the efficacy of AI and humans in diagnosing illnesses.

According to this study, AI performed better at diagnosing than humans did—in fact, when compared to less skilled doctors, AI outperformed humans in this regard [8–12]. Seventy to eighty percent of individuals with liver cirrhosis infection develop to the terminal stages of the disease because there are no symptoms, making early detection hard. Moreover, chronic sickness patients may go years without showing any symptoms. By the time it's all over, the liver's ability to operate is completely damaged, making treatment unfeasible [13]. In a paper cited in Michigan Medicine [14], Lok claims that the optimal course of treatment with the highest likelihood of recovery is only possible when the disease is discovered early on. Lok claims that individuals with liver cirrhosis frequently have symptoms during the onset of the disease, which raises the risk of liver cancer. Because of this, a novel and useful liver cirrhosis diagnostic technique satisfies the need for early identification and enables medical professionals to treat patients in a timely manner. Early identification also lessens the chance that the virus may infect other people [15]. Early diagnosis of acute infections and chronic diseases may be facilitated by AI-based disease diagnostics and prediction systems.

In a study published in 2014, Keltch, Lin, and Bayrak employed 424 liver cirrhosis patients' publicly available data to test four different types of AI algorithms. Their proposed methodology helps predict fibrosis stage by comparing the results of biopsies with the findings of standard serum markers. The healthcare system as a whole could profit from Keltch et al.'s (2014) novel approaches and other AI techniques that could help forecast hepatitis B and C in millions of patients worldwide without the need for biopsies. The authors of [16] also suggest that a variety of structured and unstructured healthcare data sets may be handled by AI techniques in their study.

Artificial intelligence technologies that are widely used include machine learning methods for structured datasets. Pietrangelo (2018) and Lok (2016) contend that early detection of the infectious stage is the only way to get the greatest treatment outcomes and minimize challenges brought on by disease progression. Therapy should thus start as soon as is practical. Lok and her colleagues are developing a novel method to increase the accuracy of predicting the chance of developing fibrosis and its progression from mild to moderate [17–20]. This method makes use of machine learning techniques to include many datasets. In [21], an AI method for identifying human sickness is built using 29 algorithmic parameters that correspond to the symptoms of liver cirrhosis infection.

A dataset containing 29 symptom characteristics from Egyptian patients receiving liver cirrhosis virus treatment for 18 months was made available by Kamal et al. (2019). Using the proposed diagnostic technique, researchers, scientists, and medical professionals will be able to predict infection stages without requiring liver samples from patients. To get our results in this study, we used a variety of well-known machine learning methods. With 86.74 percent, 82 percent, and 80.72 percent of F1-scores, respectively, Random Forest, Logistic Regression, and Gradient Boosting were the most successful algorithms. The models employed in this investigation have an accuracy percentage that is significantly greater than the accuracy percentage of the models used in previous studies, suggesting that the models used in this study are more trustworthy.

Several model comparisons have shown them to be resilient, and the scheme may be built based on the analytical results of the study. This is how the rest of the research is organized. Related work is covered in the second section; procedures and experimental techniques are covered in the third section; findings analysis and conclusions are included in the fourth and fifth sections.

By looking at a variety of characteristics, the research aims to improve outcomes by promoting early identification and timely treatment. Combining current technologies with patient knowledge might lead to amazing accomplishments in mental health treatment.

1.2 Motivation

Liver cirrhosis is a chronic and progressive liver disease that poses a significant global health burden, leading to substantial morbidity and mortality. Early detection and accurate prediction of liver cirrhosis can play a crucial role in improving patient outcomes by enabling timely interventions and personalized treatment strategies. Traditional diagnostic methods often rely on invasive procedures and may lack sensitivity, leading to missed diagnoses and delayed interventions. In recent years, machine learning techniques have emerged as powerful tools for medical diagnosis and prediction tasks. Leveraging the vast amounts of data available from various sources, including clinical records, imaging studies, and laboratory tests, machine learning models can effectively learn complex patterns and relationships to predict the risk of liver cirrhosis. This research project aims to explore the potential of machine learning algorithms in predicting liver cirrhosis risk based on diverse patient data. By harnessing the predictive capabilities of machine learning, we strive to develop accurate and clinically relevant models that can assist healthcare professionals in early identification and intervention for patients at risk of liver cirrhosis. Ultimately, this work seeks to contribute to the advancement of personalized medicine and improve patient outcomes in the management of liver cirrhosis.

1.3 Rationale of the Study

The high morbidity and death rates associated with liver cirrhosis make it a major global public health problem. Improving patient outcomes and cutting costs associated with healthcare need early identification and prompt action. Machine learning techniques offer a promising approach for accurate and efficient prediction of liver cirrhosis risk. By leveraging advanced algorithms and large datasets, this research aims to develop predictive models that can identify individuals at high risk of developing cirrhosis, enabling early intervention strategies and personalized healthcare approaches. This study addresses the critical need for innovative solutions in liver disease management, contributing to improved patient care and healthcare system efficiency.

1.4 Research Questions

1. Which Algorithm for Machine Learning Provides the Best Predictive Results?

2. Can Machine Learning Models Improve Predictive Accuracy Compared to Traditional Research Approaches?

3. Using Machine Learning to Predict Liver cirrhosis, Which Ethical Issues Should Be Considered?

1.5 Expected Output

- By using specific traits and datasets, this study attempts to create a machine-learning model that can accurately predict the beginning of liver cirrhosis.
- Additionally, it has the ability to assess and contrast the machine learning model's performance with that of conventional diagnostic techniques.
- It can also highlight the potential improvements in accuracy and early detection offered by the predictive model. There can be a comparison of the effectiveness of different machine learning algorithms, providing clear sightedness into which algorithms perform best in predicting Liver cirrhosis and under what status or condition.
- This study will help advance the understanding of Liver cirrhosis, emphasizing how the research findings may inform future studies and enhance clinical approaches to mental health prediction, management and analysis.

1.6 Project Management and Finance

To properly manage a research project, I followed a systematic and organized approach. Here are the key steps I took:

- Established research objectives based on SMART criteria, ensuring specificity, measurability, achievability, relevance, and time-bound focus.
- Created a detailed project plan outlining tasks, milestones, timelines, and resources for research, defining scope and allocation.
- Reviewed literature to identify knowledge gaps, potential methodologies, and design.
- Designed research methodology considering data collection, ethical considerations, sample size, statistical methods and developing mobile app.
- Sought ethical approval from relevant institutions for research compliance.
- Conducted data collection, ensure documentation, quality control, analyze using statistical methods.
- Interpreted findings, communicate findings to stakeholders.
- Monitored progress, identify deviations, adjust plan, review, revise for research objectives.
- Documented research activities for transparency, reproducibility, and compliance.

This project was completely self-funded, with no outside funding. This initiative received no financial aid or funding.

1.7 Report Layout

The remainder of the thesis is organized as follows:

Chapter 2: This section provides an overview of existing literature and discusses various approaches to sentiment and emotion classification. It also outlines the scope, significance, and challenges within this domain.

Chapter 3: Building upon prior research, this chapter identifies challenges and details the procedure of the proposed model, emphasizing strategies to overcome these challenges. It covers aspects such as algorithmic approaches, feature selection, data annotation, data preparation, and collection methods.

Chapter 4: This chapter presents the final outcomes of the proposed tools for multilabel Liver cirrhosis prediction. It includes a comprehensive discussion of the overall process.

Chapter 5: The societal impact, importance, and sustainability of this research are deliberated in this section.

Chapter 6: Concluding the thesis, this section provides a concise summary of the research, followed by an exploration of its limitations. Additionally, recommendations and potential avenues for further research are discussed.

CHAPTER 2

BACKGROUND STUDY

2.1 Preliminaries

A comprehensive research strategy is defined in the study's preliminary part on the use of machine learning to predict liver cirrhosis. The headline of the research introduces it and gives key background information. The Abstract that follows provides a brief summary of the study's goals, procedures, results, and conclusions. The References honor writers, which fosters a spirit of cooperation. Readers can locate information more easily with the help of the report's Table of Contents, List of Figures and Tables, and List of Symbols and Terms. Clarity is ensured by an index, which defines targeted words. As the Introduction said, machine learning-based liver cirrhosis prediction is significant. The Literature Review provides a thorough study of prior studies in order to identify research gaps. The section on research aims and predictions provides an explanation of the study's goals, which align with broader goals for enhancing mental health. The approach includes a detailed description of the data gathering process, research design, and ethical issues.

2.2 Related Works

The literature related to Liver cirrhosis identification, socio-demographic factors, psychological factors, and machine learning applications in mental health. I have studied some recent research works and the analysis of those studies are given below:

Sreejith et al. [3] In their paper, the researchers also evaluated the classification performance by using ILPD, Thoracic Surgery dataset, and Pima Indian Diabetes dataset. The main aim of this study is one: to compare performance before and after feature selection in addition to using class balancing SMOTE technique and CMVO-based feature selection using AR feature extractor and random forest classifier. On the ILPD dataset, they obtained 69.43% accuracy without OSMOTE and CMVO-based feature selection, 82.620% accuracy with OSMOTE and CMVO-free selection, and 82.46% accuracy with OSMOTE and CMVO-free selection.

Kuzhippallil et al. [22] compared and enhanced the performance of datasets for the classification of chronic liver disease using a multi-strategy that used a variety of data preprocessing techniques and feature selection techniques, like imputation for missing values, isolation forest detection and elimination of outliers, removal of duplicate values, and so forth. Following the suggested method—which includes MLP, KNN, LR, DT, RF, Gradient Boosting, AdaBoost, XGBoost, Light GBN, and Stacking Estimator—they categorized liver patients and compared the accuracy, which came out to be 82%, 79%, 76%, 84%, 88%, 83%,86%, 86%, and 85%, respectively.

Rahman et al. [23] evaluated six distinct classification methods, including Random Forest (RF), KNN, Decision Tree (DT), Support Vector Machines (SVM), NB, and Logistic Regression (LR), to forecast liver disorders. LR had the greatest accuracy of 75%, according to their comparison research.

Arshad, Insha, et al. [24] Digestion on the discovery of liver disease because of exorbitant alcoholism through data mining techniques was done. The rule proposed distinguishing just as to anticipate the nearness of liver disease using data mining calculations. After showing the rules, they use different data mining calculations to train and test the dataset towards recognition of the liver disease. This considers the information that was gotten from UCI repository dependent on lab results and contains 7 distinct characteristics with 345 occurrences.

Arshad, Insha, et al. [25] There were five possible arrangement calculations for the diagnosis of liver disease. In this project, two different types of liver informational indexes—BUPA liver issue information and India Liver Patient Information (ILPD)—have been subjected to calculations using Naive Bayes grouping, C 4.5 Decision Tree, Back Propagation, K-Nearest Neighbor, and Bolster Vector Machine algorithms. In this project, more computations are used to evaluate the grouping execution in terms of accuracy, precision, sensitivity, and specificity when characterizing the dataset of liver patients. Finally, they have achieved the highest level of accuracy with the use of Support Vector Machine and K-Nearest Neighbor computations.

Gregory (2015). [26] Two real liver patient datasets had been suggested, and classification models were created to predict the diagnosis of liver. Both the suggested datasets have been exposed to the 11 data mining classification algorithms, and the overall performance of all the classifier results are considered in terms of accuracy, precision, and recall. Based on the obtained outcomes, the classification accuracy is fair when employing FT Tree algorithm compared to other solutions, and it exhibits the high performance based on the attributes, i.e., 78.0% of Accuracy, 77.5% of Precision, 86.4% Sensitivity, and 38.2% of Specificity.

Rahman, A. Sazzadur, et al. [27] This research explores the application of Machine Learning (ML) algorithms for predicting Chronic Liver Disease (CLD), aiming to mitigate the high diagnostic costs associated with this global health challenge. Six ML techniques—Logistic Regression, K Nearest Neighbors, Decision Tree, Support Vector Machine, Naïve Bayes, and Random Forest—are evaluated based on accuracy, precision, and other performance metrics using a dataset of 583 liver patient records. Logistic Regression emerges as the most effective classifier, achieving a 75% accuracy rate. This study underscores the potential of ML in enhancing CLD prediction, offering valuable insights for healthcare decision support systems. Further research avenues include exploring additional ML algorithms to improve prediction accuracy.

Jamila, Ganty, et al. [28] This literature review highlights the significance of machine learning in predicting liver cirrhosis, a prevalent chronic liver disease. Early detection is crucial for effective treatment and prevention of complications. Studies employ algorithms like Naïve Bayes, CART, and SVM to develop predictive models based on patient data, with SVM showing superior performance. Evaluation metrics such as accuracy, precision, recall, and F1 Score assess model performance. These models aid healthcare providers in identifying at-risk individuals and initiating timely interventions. Future research may focus on refining algorithms and validating model performance in diverse populations. Overall, machine learning holds promise in improving liver cirrhosis diagnosis and management, enhancing patient outcomes.

Wu, Chieh-Chen, et al. [29] carried out a prediction study on Fatty Liver Disease (FLD) patients. A total of 700 patient records with screening tests for fatty liver disease were gathered by the study team from New Taipei Hospital; of these, 577 records were taken into consideration based on the patient's age and enough data. 377 of the 577 individuals had fatty liver disease, whereas the other patients did not. Age, gender, systolic and diastolic blood pressure, belly circumference, glucose level, triglyceride, HDL-C, SGOT-AST, and SGPT-ALT are among the patient health characteristics included in the dataset. Normalization was carried out together with the application of the Synthetic Minority Over-Sampling Technique (SMOTE) during the data preparation phase. The following phase involved using four machine learning algorithms: Random Forest, Naïve Bayes, Artificial Neural Network, and Logistic Regression with 3, 5, and 10-fold cross-validation. Apart from the accuracy, each algorithm's area under the receiver operating curve was noted. Out of all the findings, Random Forest had the highest accuracy across all cross-validations.

Singh et al. (2020) [30] concentrated their effort on developing software for simple prediction and utilizing various feature selection and classification techniques to forecast liver disease. The Indian Liver Patient Records dataset was used for the investigation. Using WEKA's Correlation-based Feature Selection Subset Evaluator and the Greedy Stepwise search approach, a few characteristics were eliminated during the feature selection stage. Using this procedure, just five (5) qualities were chosen: aspartate aminotransferase, alkaline phosphatase, alamine aminotransferase, total bilirubin, and direct bilirubin. This led to the use of six distinct classification techniques: Instant-based Classification (IBk), Random Forest, Sequential Minimal Optimization (SMO), Naïve Bayes, and Logistic Regression. Of these, Logistic Regression yielded the greatest accuracy, at 74.36%. Naïve Bayes produced the least accuracy (55.9%).

Bhupathi, Deepika, et al. [31] This paper investigates the effectiveness of machine learning algorithms in predicting liver diseases, aiming to develop a liver disease prediction (LDP) method applicable to healthcare professionals and researchers. Five algorithms - SVM, Naïve Bayes, K-NN, LDA, and CART - are compared using R and Python. The study follows the SEMMA lifecycle, including data sampling, exploration, modification, preprocessing, modeling, assessment, and result analysis. Experimental results show K-NN and autoencoder networks achieving the highest accuracies of 91.7% and 92.1%, respectively. The research highlights the importance of feature selection and dataset expansion to improve prediction accuracy and underscores the potential of machine learning in enhancing healthcare outcomes through early disease detection.

Elias, et al. [32] The paper investigates liver disease prediction using machine learning (ML) techniques, emphasizing the urgency of early detection for effective medical intervention. It reviews the liver's critical functions, various liver diseases, and their causes. Leveraging the Indian Liver Patients' Records dataset, the study employs data preprocessing, feature analysis, and evaluates multiple ML models. Notably, the Voting classifier emerges as the most effective, surpassing other models in accuracy and predictive performance. The findings highlight the significance of ML in healthcare and underscore the potential of the proposed model for liver disease prediction. Despite certain limitations, such as dataset constraints, the study provides valuable insights and sets the stage for future research aimed at refining predictive models and enhancing early disease detection in liver health management.

Gogi, Vyshali J., et al. [33] This paper explores the application of data mining classification techniques in predicting liver disease, utilizing Decision Tree, Linear Discriminant, SVM Fine Gaussian, and Logistic Regression algorithms. Leveraging MATLAB2016 for implementation, the study analyzes laboratory parameters of patients to ascertain the most effective algorithm. Results indicate Linear Discriminant algorithm achieves a prediction accuracy of 95.8% with an ROC of 0.93. The findings underscore data mining's potential in healthcare diagnostics, particularly in liver disorder prediction. By highlighting the efficacy of computational methods in medical decision-making, the research contributes to advancing healthcare analytics. Moving forward, further investigations are warranted to optimize predictive modeling and algorithmic performance, ultimately enhancing medical knowledge and patient care in the ever-evolving landscape of healthcare analytics.

Hanif, Ishtiaqe, et al. [34] This study investigates the efficacy of machine learning algorithms in the early diagnosis of liver cirrhosis, a challenging blood-borne illness often asymptomatic in its initial stages. Three distinct models utilizing Support Vector Machine, Decision Tree Classification, and Random Forest Classification were developed and evaluated. Among these, the Random Forest algorithm demonstrated superior performance, achieving an accuracy of approximately 97 percent. Leveraging the Liver Cirrhosis dataset, the study showcases improved predictive accuracy compared to previous research, validating the reliability of the models employed. Comparative analyses and evaluation metrics such as precision, recall, and F1-score underscore the robustness of the proposed AI-driven diagnostic approach. The findings highlight the potential of machine learning in enhancing early detection and management of liver cirrhosis, with implications for improving patient outcomes and public health initiatives.

Li, Xiaoguo, et al. et al. [35] The current developments in artificial intelligence for the noninvasive detection of gastric varices and portal hypertension, as well as the tracking of risk assessment for associated consequences in clinical practice, are the main topics of this study. To predict liver cirrhosis, the study uses three distinct machine learning models: Random Forest Classification, Decision Tree Classification, and Support Vector Machine. The models' accuracy, recall, and F1-score are used to assess their performance. With an accuracy of almost 97%, the data indicate that Random Forest was the best-performing algorithm. Additionally, the study uses Pearson Correlation Coefficient based feature selection (PCC-FS) to remove elements from the dataset that aren't relevant. AdaBoost is a boosting technique that is used to improve such algorithms' predicting performance. Evaluation criteria for the comparison analysis include accuracy, ROC, F-1 score, precision, and recall. The usage of artificial intelligence technologies, according to the paper's conclusion, has the potential to completely change our profession by utilizing vast quantities of data to individually tailor care to the appropriate patient at the right time. The limited sample size and the absence of external model validation are two of the study's shortcomings. Subsequent research endeavors ought to concentrate on verifying the models using more extensive datasets and contrasting their efficacy with that of alternative machine learning techniques.

2.3 Comparative Analysis and Summary

The study uses a wide range of machine learning (ML) algorithms and evaluates and contrasts them in order to determine the advantages and disadvantages of each algorithm for Liver cirrhosis prediction. Performance metrics including as recall, accuracy, precision, and F1 score are used to compare and contrast the efficacy of these algorithms. The research project on predicting Liver Cirrhosis using machine learning algorithms has shown that the Random Forest model guarantees an accuracy level of 86.74%. Based on the comparison with other algorithms, including Decision Trees, Gradient Boosting, Logistic Regression, and Support Vector Machines, RF demonstrates the highest accuracy. Therefore, RF is more effective than its competitors in a high-quality forecast, focusing on the complex dataset and intricate interrelations of patterns coinciding with the development of Liver Cirrhosis. Being based on the principles of ensemble learning, such a model shows acceptable performance, proving its effectiveness in early detection and forecasting of Liver Cirrhosis. In summary, this thorough method highlights the possibility for early Liver Cirrhosis detection and intervention, provides helping benefits to health outcomes, and guides the selection of models for better accuracy.

2.4 Scope of the Problem

The possible usages for machine learning for Liver cirrhosis prediction are huge, given the condition's prevalence and global impact. Machine learning for prediction of liver cirrhosis involves designing models that will analyze patients' data and predict whether they are likely to get the liver disease. The topic includes the different categories of machine algorithms, feature selection methods, and model evaluation to develop robust and accurate predictive models. On the other hand, the subject incorporates the analysis of datasets that contain clinical data, genetic data, and patients' biomarkers, among other demographic information to increase the predictive capacity of the models. Moreover, the research will identify some of the challenges such as class imbalance, feature selection, and understanding of the black-box models which inhibit the implementation of such models in clinical practice.

2.5 Challenges

Developing machine learning models for liver cirrhosis prediction presents several challenges:

- Data Imbalance: Addressing the imbalance between cirrhotic and non-cirrhotic patients in the dataset to prevent biased predictions.
- Feature Selection: Identifying the most relevant features from complex medical data while ensuring model interpretability.
- Data Quality: Handling missing values, outliers, and noisy data to improve model performance.
- Generalization: Ensuring that predictive models generalize well across diverse patient populations and healthcare settings.
- Clinical Relevance: Incorporating domain knowledge and ethical considerations to ensure the practical utility and ethical soundness of predictive models in clinical practice.

Addressing these challenges requires interdisciplinary collaboration and rigorous validation methodologies to develop accurate and clinically relevant predictive models for liver cirrhosis.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research Subject and Instrumentation

The study population growth consists of people who have been diagnosed with liver cirrhosis or who are showing symptoms consistent with the illness. This research project is relevant to liver cirrhosis prediction using machine learning algorithms. The main purpose of this study to build predictive models, which can forecast the disease state. It means identifying patients under risk of liver cirrhosis developing possibility according to the clinical and demographic characteristics. The instrumentation in this study would be the advanced machine learning algorithms, such as Support Vector Machines, Gradient Boosting Classifier, Random Forest Classifier, Decision Tree Classifier, and Logistic Regression. They are used in the dataset that consists of all the essential medical data. The result is several strong predictive models that can help recognize liver cirrhosis as soon as possible. The testing process is guided by principles of ethics that guarantee participant privacy and confidentiality. This instrumentation improves the study's ability to understand the complex issues of liver cirrhosis prediction by combining conventional assessments with modern machine learning algorithms.

3.2 Data Collection Procedure

For my research on liver cirrhosis prediction, I obtained data from two distinct sets accessible on Kaggle. These datasets include various variables associated with liver health, such as demographics, past history and comorbidities, laboratory investigations, and imaging. Once we obtained access to the datasets, we vetted and clean the data to ensure consistency and robustness. I proceeded to combine the datasets using common key identifiers to develop our final merged dataset, which we use to train and test our machine learning models.

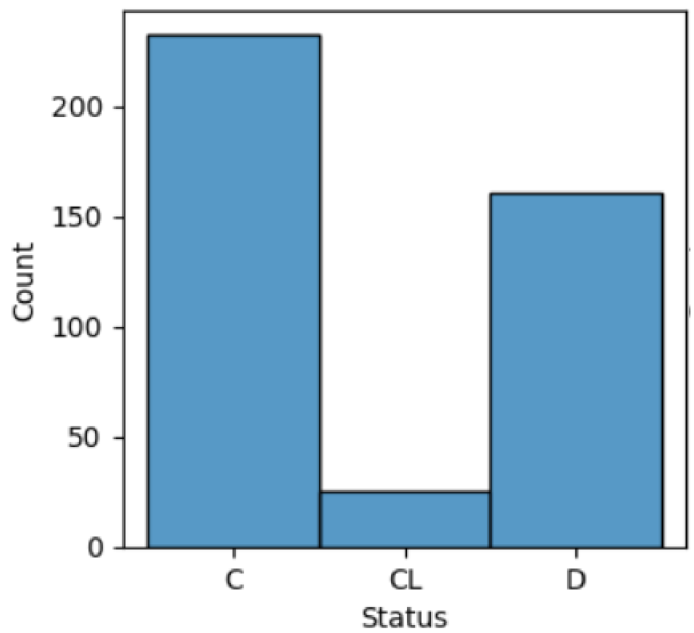

Figure 3.1: Dataset liver cirrhosis Status Bar Plot

The above bar plot of figure 3.2 represents the distribution of three categorical attributes, "C", "CL" and "D," with corresponding counts of 232, 25 and 161, respectively. The horizontal axis typically represents the categories (in this case, "C", "CL" and "D,"), while the vertical axis represents the count or frequency. Three bars are present on the plot—one for "C" with a height of 232 and another for "D" with a height of 161. The bars visually illustrate the comparison between the three attributes, showing that the count for "C" is slightly higher than "CL" and "D".

Dataset

Out of the 23 columns in my dataset, 22 features are present. There is an additional column designated for status. Age, gender, and 20 questions are among the features.

ID	N_Days	Status	Drug	Age	Sex	Ascites	Hepatome	Spiders	Edema	Bilirubin	Cholesterc	Albumin	Copper	Alk_Phos	SGOT	Tryglicerid	Platelets	Prothromb	Stage
1	400	D	D-penicilla	21464	F	Y	Y	Y	Y	14.5	261	2.6	156	1718	137.95	172	190	12.2	4
2	4500	C	D-penicilla	20617	F	N	Y	Y	N	1.1	302	4.14	54	7394.8	113.52	88	221	10.6	3
3	1012	D	D-penicilla	25594	M	N	N	N	S	1.4	176	3.48	210	516	96.1	55	151	12	4
4	1925	D	D-penicilla	19994	F	N	Y	Y	S	1.8	244	2.54	64	6121.8	60.63	92	183	10.3	4
5	1504	CL	Placebo	13918	F	N	Y	Y	N	3.4	279	3.53	143	671	113.15	72	136	10.9	3
6	2503	D	Placebo	24201	F	N	Y	N	N	0.8	248	3.98	50	944	93	63	NA	11	3
7	1832	C	Placebo	20284	F	N	Y	N	N	1	322	4.09	52	824	60.45	213	204	9.7	3
8	2466	D	Placebo	19379	F	N	N	N	N	0.3	280	4	52	4651.2	28.38	189	373	11	3
9	2400	D	D-penicilla	15526	F	N	N	Y	N	3.2	562	3.08	79	2276	144.15	88	251	11	2
10	51	D	Placebo	25772	F	Y	N	Y	Y	12.6	200	2.74	140	918	147.25	143	302	11.5	4
11	3762	D	Placebo	19619	F	N	Y	Y	N	1.4	259	4.16	46	1104	79.05	79	258	12	4
12	304	D	Placebo	21600	F	N	N	Y	N	3.6	236	3.52	94	591	82.15	95	71	13.6	4
13	3577	C	Placebo	16688	F	N	N	N	N	0.7	281	3.85	40	1181	88.35	130	244	10.6	3
14	1217	D	Placebo	20535	M	Y	Y	N	Y	0.8	NA	2.27	43	728	71	NA	156	11	4
15	3584	D	D-penicilla	23612	F	N	N	N	N	0.8	231	3.87	173	9009.8	127.71	96	295	11	3

Figure 3.2: Dataset

3.3. Data Analysis

From my data analysis, I obtained promising results in the prediction of liver cirrhosis through machine learning. Based on our assessment, the accurate percentage was exceptionally high, indicating that the prediction was quite effective. Furthermore, other performance factors, including the precision, recall, and F1-score, highly contributed to understanding how well it could predict. In addition to the abovementioned factors, feature importance scores can also be considered, which helped to determine which factors were more crucial in predicting whether the patient had liver cirrhosis or not.

3.4 Proposed Methodology

our research aims to utilize powerful artificial intelligence technology, machine learning, for liver cirrhosis prediction. Firstly, we collect extensive clinical data and preprocess them carefully. After data cleaning, we use feature selection to select significant indicators. We explore multiple classification algorithms, such as logistic regression, decision trees, and gradient boosting, to build a suitable model. Then, after intensive training and evaluation, we optimize the model, focusing on evaluation metrics such as accuracy and area under the ROC curve. Finally, we validate the model on independent datasets to ensure its reliability and generalizability. Our ultimate goal is to create an accurate and interpretable prediction model which can assist in early discovery and prognosis of liver cirrhosis, improving patient care and human resources while also simplifying medical decision-making. Here is a general summary in the below flowchart in Figure 3.4:

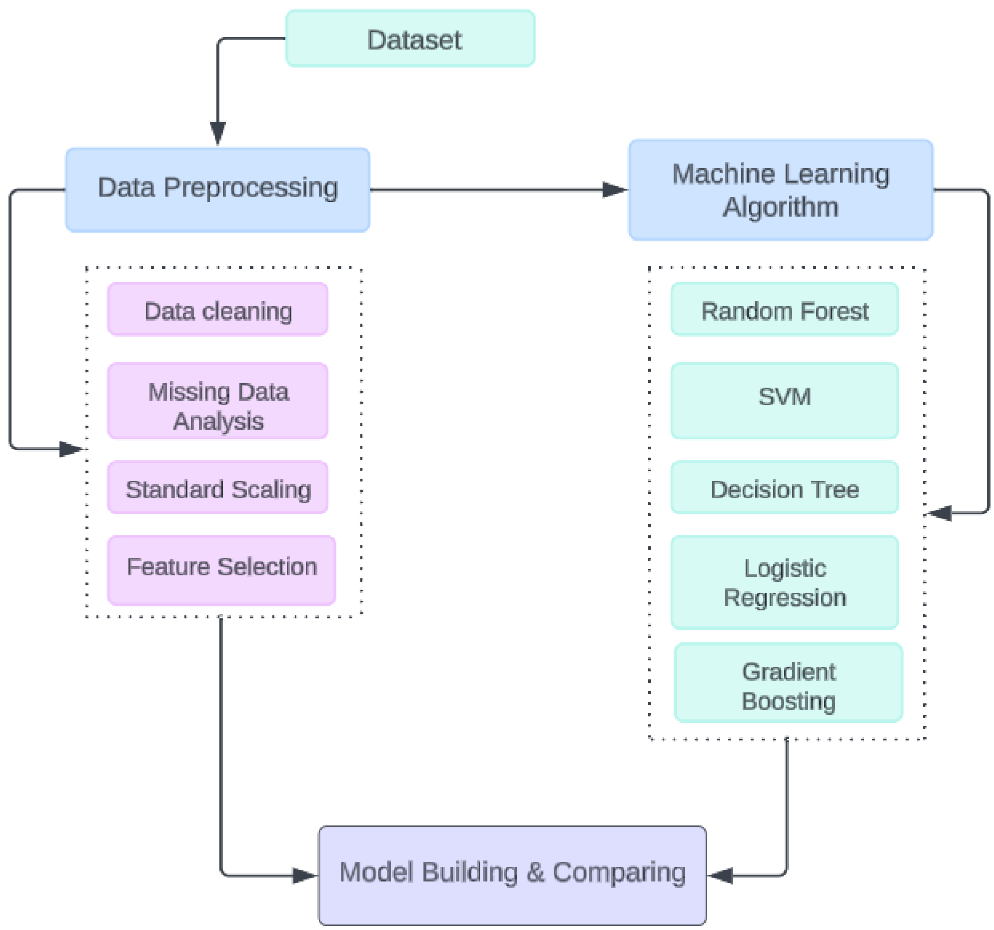

Figure 3.3: Methodology Flowchart

Missing Value Removal

To ensure the dataset's integrity, missing values were found and carefully removed. Depending on the nature and quantity of missing data, several procedures, such as estimation or removal of partial information, were used. This procedure looked to improve the quality and completeness of the dataset for analysis that followed.

Label Encoding

To ease machine learning model training, categorical variables were converted to numerical representation. To convert categorical attributes such as gender and other information into a format appropriate for input into machine learning algorithms, techniques such as one-hot encoding were used.

Model Selection

Selecting the right models to train data is crucial for machine learning predictions of bipolar disorder. The models that can be taken into consideration are Decision Tree, Random Forest, Gradient Boosting, Support Vector Machines (SVM) and Logistic Regression. LR is a straightforward yet effective binary classification technique. One effective tool that can assist boost machine learning model performance is the voting classifier. To produce a prediction that is more accurate, they combine the results of numerous distinct models that were each trained using a different algorithm.

Support Vector Machine

Support Vector Machine (SVM) is a sophisticated classification and regression technique. It finds a decision boundary that optimizes the margin between classes, resulting in improved generalization. SVM can handle non-linear interactions by employing kernel functions to change the data. It is resistant to outliers and excels in high-dimensional spaces. Regularization aids in the control of overfitting. SVM is adaptable, as it can do binary and multi-class classification as well as regression. It does, however, have lengthier training and prediction periods and necessitates careful parameter selection.

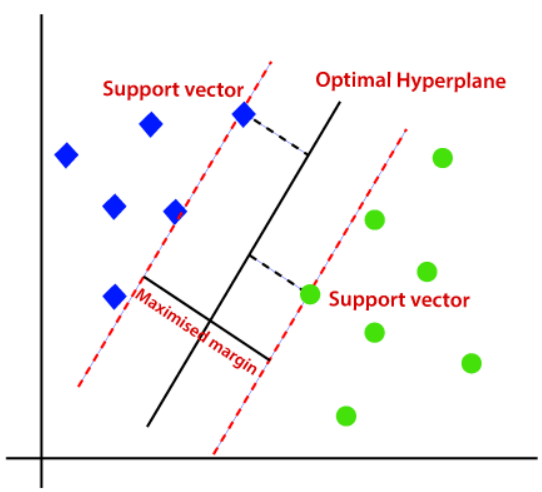

Figure 3.4: SVM Model Architecture

Logistic Regression

A popular and essential statistical model for binary classification is logistic regression. It illustrates the connection between the input variables and the probability of falling into a particular class, as opposed to linear regression. In logistic regression, the linear combination of input features is converted to a probability score using the logistic function (sigmoid). This probabilistic technique enables intuitive interpretation as well as the estimation of class probabilities. It is noise-resistant and computationally efficient, and it works well with both continuous and categorical predictors.

Linear Combination:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

Here, z represents the linear combination of input features, β_0 is the intercept term, β_1, β_2, ..., β_p are the coefficients for each input feature (x_1, x_2, ..., x_p).

Sigmoid Transformation:

$$p = 1 / (1 + e^{(-z)})$$

The sigmoid function (1 / (1 + e^(-z))) maps the linear combination (z) to a probability score (p) between 0 and 1.

Figure 3.5: LR Model Architecture

Gradient Boosting (GB)

Gradient Boosting is an ensemble learning technique that combines the predictions of multiple weak learners, most commonly decision trees, to produce a strong predictive model. Gradient Boosting, unlike Random Forests, builds trees sequentially, with each tree correcting the errors of its predecessor. On the basis of the original data, a simple model, often a shallow decision tree, is built. Following trees are built to correct the errors introduced by the combined predictions of the existing trees. The main idea is to fit each new tree to the ensemble's residuals (the differences between the actual and predicted values). Each tree is trained to capture patterns that the previous trees were unable to adequately model. Gradient Boosting term "Gradient" refers to the use of gradient descent optimization to minimize the loss function, guiding the construction of subsequent trees in the direction that minimizes errors. The learning rate parameter governs each tree's contribution to the ensemble. Gradient Boosting is more sensitive to overfitting than

Random Forests, but it often achieves higher predictive accuracy when properly tuned. It is widely used in machine learning applications such as regression and classification.

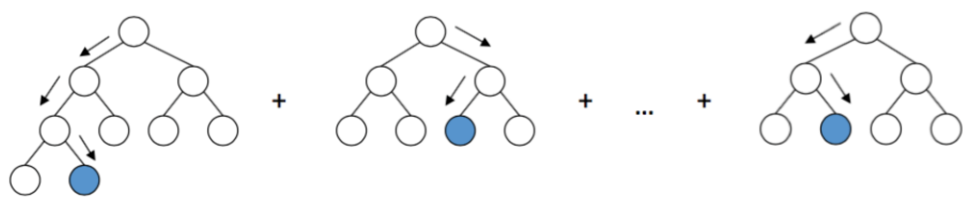

Figure 3.6: Gradient Boosting (GB) Architecture

Decision Tree

Decision trees are machine learning algorithms that are adaptable and interpretable and are used for classification and regression applications. They can handle both categorical and numerical data, making them extremely versatile. Because each branch indicates a choice based on a feature, decision trees create transparent models that aid in human comprehension. They can manage imbalanced datasets and are resistant to noisy and missing data. Decision trees, on the other hand, can overfit, catching noise or irrelevant patterns. Pruning and ensemble approaches can help with this problem. Large or high-dimensional datasets can be difficult to handle, although ensemble methods such as random forests or boosting can improve performance. Decision trees provide a good blend of interpretability and performance, making them useful in a variety of fields.

$$Entropy(s)= -P(yes)log2\ P(yes)- P(no)\ log2\ P(no)$$

$$Information\ Gain\ =\ Entropy(before) - \sum_{j=1}^{K} Entropy(j,\ after)$$

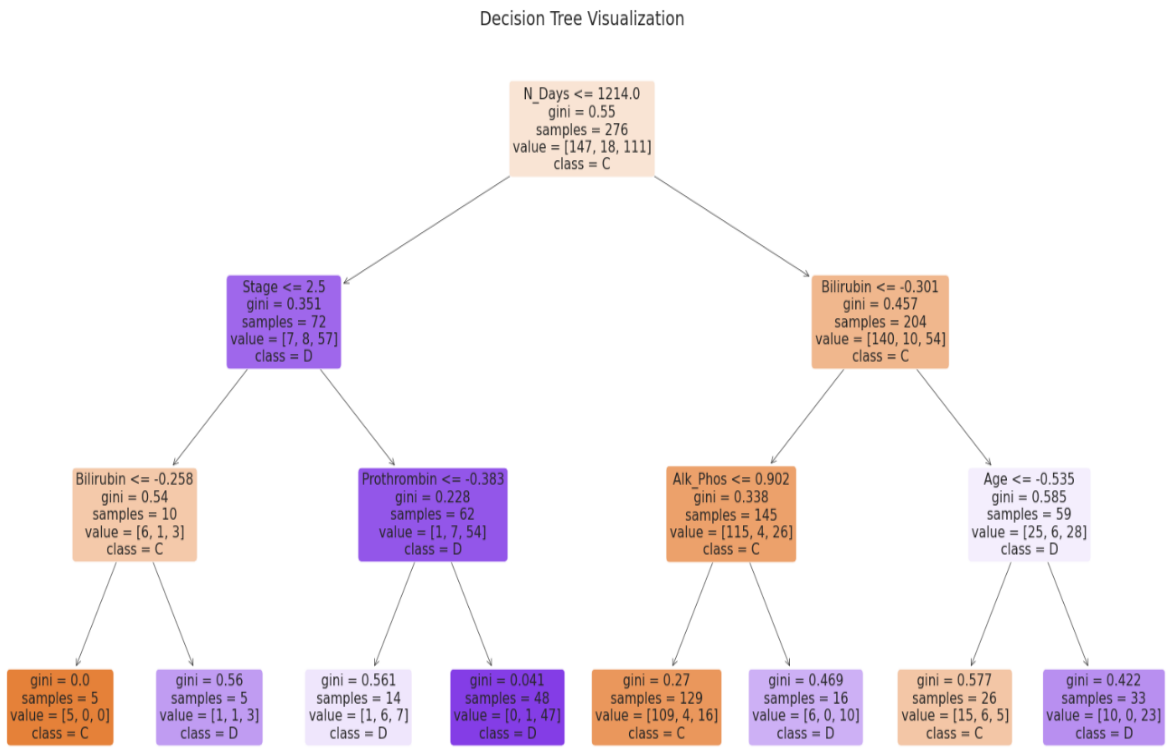

Figure 3.7: Decision Tree Architecture

Random Forest

The RF classifier is an ensemble method that uses bootstrapping and aggregation, or bagging, to train several decision trees simultaneously. Bootstrapping is the process of training several decision trees simultaneously on various subsets of the training dataset with different subsets of accessible characteristics. By guaranteeing that each decision tree within the random forest is unique, bootstrapping reduces the total variance of the RF classifier. The RF classifier has great generalization since it combines the decisions made by individual trees to get a final result. The RF classifier usually outperforms most other classification methods in terms of accuracy when overfitting issues are not present. Unlike DT classifiers, RF classifiers do not require feature scaling. Compared to the DT classifier, the RF classifier is more robust against noise in the training dataset and in the process of selecting training samples. The RF classifier is harder to grasp than the DT classifier, but it's easier to tweak the hyperparameter.

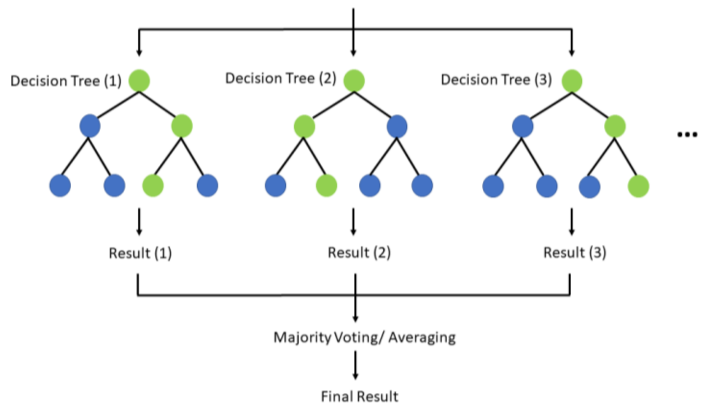

Figure 3.8: RF Architecture

3.5 Implementation Requirements

This research doesn't demand sophisticated devices for data collection and the employed methodology. Nevertheless, there is a baseline requirement for device specifications to effectively execute scraping algorithms and associated methods. The critical specifications include the CPU and RAM, as elaborated in Table 3.3. The device specifications for this study are also provided in the table. Adhering to these minimum requirements is crucial, particularly when implementing deep learning approaches. It is noteworthy that Google Colab has been utilized to run resource-intensive machine learning algorithms as well.

TABLE 3.1: IMPLEMENTATION REQUIREMENTS

Terminology	**Used in this Research**	**Minimum Requirement**
CPU	Intel® Core™ i5-8265U CPU @ 1.60GHZ	Intel® Core™ i3-9100E, AMD Ryzen™ 3 2200G, or higher)
RAM	8 GB	8 GB

.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental Setup

Google Collab was used for the training process. Management and sharing of notebooks are made simpler by the integration with Google Drive. Performance for jobs that need a lot of processing is enhanced by acceleration from GPUs and TPUs. Additionally, it excels at carrying out all machine learning tasks. In my study on identifying liver cirrhosis using machine learning, I meticulously curated a raw dataset of 418 rows and 20 columns, with the target attribute classified as 'C,' 'CL, and 'D.'

4.2 Experimental Results & Analysis

It was necessary to assess the performance of the existing models at this time. I may utilize a range of performance evaluation techniques to ensure that my proposed model performs as intended. These algorithms use previously unobserved data to assess overall performance. My approach makes it possible to create precise machine learning models for the early and accurate detection of liver cirrhosis by guaranteeing a thorough representation of important components. I started by using the dataset I had selected. I calculated the numbers that were incorrect or missing and took them out of my dataset. I put a few methods into practice and assessed how well they worked. I evaluated the Accuracy, Precision, Recall, and F-1 Score of the Confusion Matrices for my proposed methods.

Accuracy: The confusion matrix is used to compute accuracy, an intuitive performance metric that shows the percentage of correctly predicted data to all of the dataset's data. The Equation is an illustration of the accuracy formula. When the dataset's false positive and false negative values are close to equal, greater accuracy can be achieved.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Precision: In the context of these algorithms, precision is defined as the ratio of correctly predicted positive values to the total number of cases that the models expected to be positive. The following equation shows the precision formula.

$$Precision = \frac{TP}{TP + FP}$$

Recall: The percentage of correctly predicting positive values to all instances of the actual positive class in the confusion matrix is called recall. The formula for calculating recall is presented in the equation below.

$$Recall = \frac{TP}{TP + FN}$$

F1 score: The weighted average of Precision and Recall, including for both false positive and false negative values, is known as the F1 Score. The formula for the F1 score is shown in the equation below. While accuracy may not be straightforward to interpret, F1 is generally considered more informative and useful.

$$F - 1\,Score = 2 * \frac{Recall * Precision}{Recall + Precision}$$

The result of machine learning model is compared on the basis of Accuracy, Precision, Recall, F1 Score in below table of 4.1:

TABLE 4.1: PERFORMANCE EVALUATION

Model Name	Accuracy	Precision	Recall	F1-Score
RF	86.74%	87%	87%	87%
DT	69%	75%	69%	72%
GB	81.92%	80%	81%	80%
LR	82%	83%	82%	82%
SVM	60%	53%	52%	45%

Accuracy

The result study looks at the train and test accuracy and analyzes which algorithm performs best. For comparison I have applied popular machine learning algorithms to check which performs perfectly. However, the RF gave the highest accuracy of 86.74%. The figure 4.1 shows the accuracy comparison of the different model:

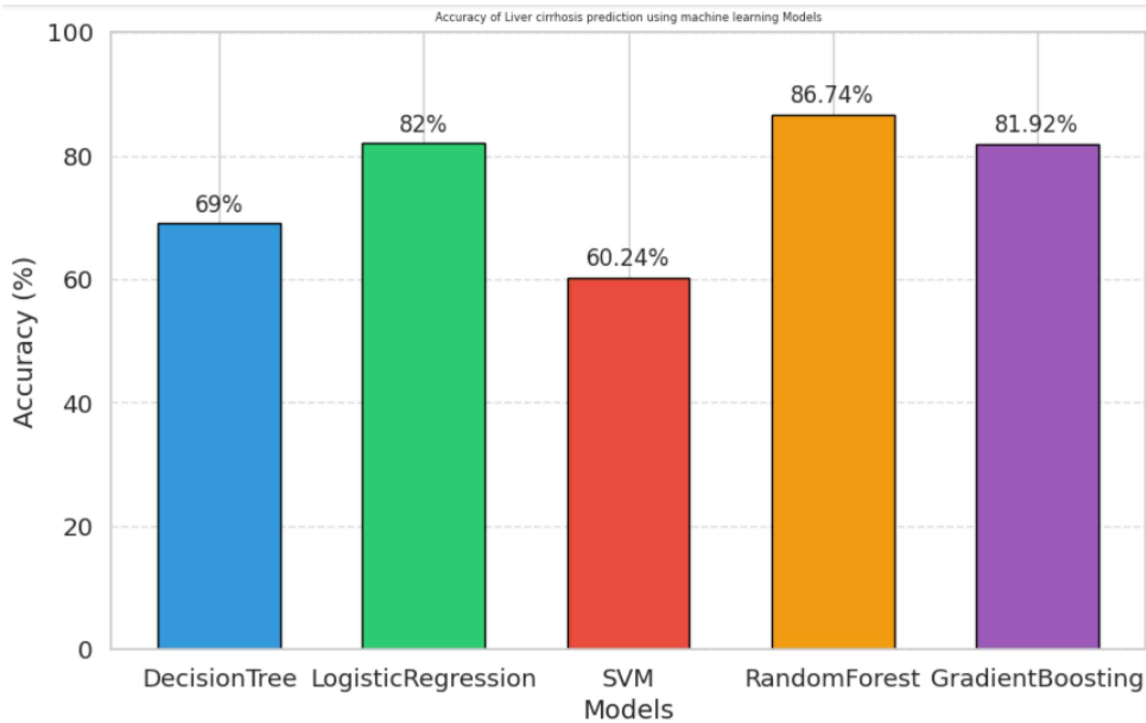

Figure 4.1: Accuracy Comparison of Machine Learning Models

The algorithmic performance of Random Forest (RF) was used to achieve the highest accuracy of 86.74%, SVM got 60.24%, GB got 81.92%, DT achieved 69% and LR got 82%.

Decision Tree:

Achieved the accuracy of 69%, Precision score of 75%, Recall score of 69% and F1-score of 72%. Below at table 4.3 I have performance evaluation of DT:

TABLE 4.2: PERFORMANCE EVALUATION(DT)

	Precision	Recall	F1-Score	Support

C	0.80	0.72	0.76	50
CL	0.00	0.00	0.00	1
D	0.66	0.66	0.66	32
Accuracy			0.64	83
Macro avg	0.49	0.46	0.47	83
Weighted avg	0.73	0.69	0.71	83

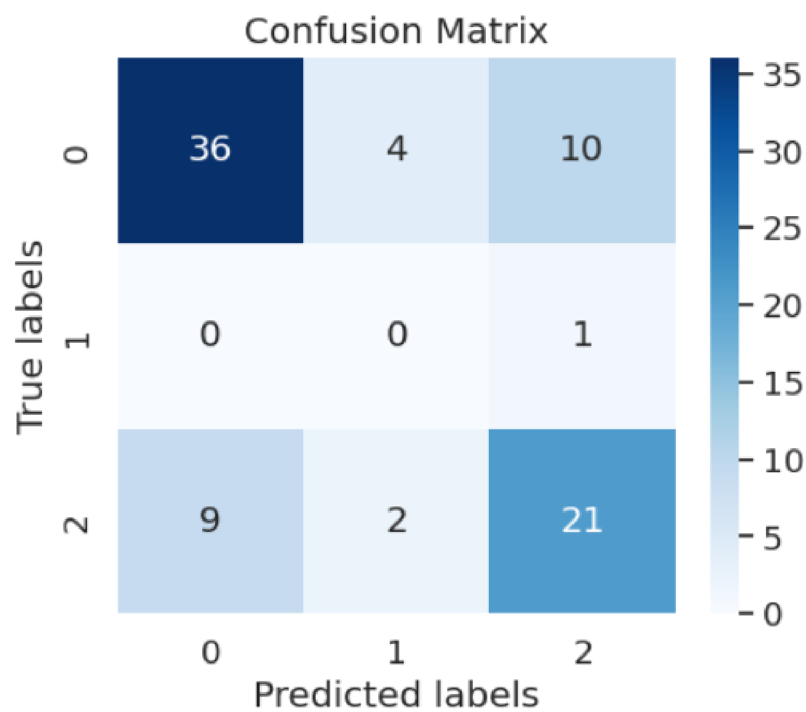

Figure 4.2: Confusion Matrix DT

Gradient Boosting:

Achieved the accuracy of 82%, Precision score of 80%, Recall score of 81% and F1-score of 80%. Below at table 4.5 I have performance evaluation of GB:

TABLE 4.3: PERFORMANCE EVALUATION(GB)

	Precision	Recall	F1-Score	Support
C	0.88	0.84	0.86	50
CL	0.00	0.00	0.00	1
D	0.76	0.81	0.79	32
Accuracy			0.82	83
Macro avg	0.55	0.55	0.55	83
Weighted avg	0.82	0.82	0.82	83

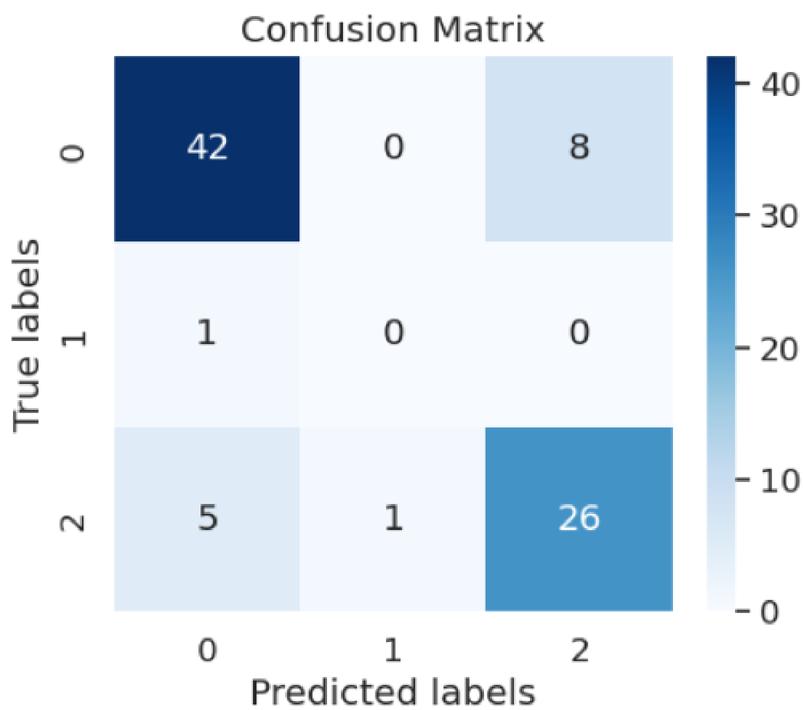

Figure 4.3: Confusion Matrix GB

Random Forest:

Achieved the accuracy of 86.74%, Precision score of 87%, Recall score of 87% and F1-score of 87%. Below at table 4.6 I have performance evaluation of RF:

TABLE 4.4: PERFORMANCE EVALUATION(RF)

	Precision	Recall	F1-Score	Support
C	0.87	0.92	0.89	50
CL	0.00	0.00	0.00	1
D	0.87	0.81	0.84	32
Accuracy			86.74	83
Macro avg	0.58	0.58	0.58	83
Weighted avg	0.86	0.87	0.86	83

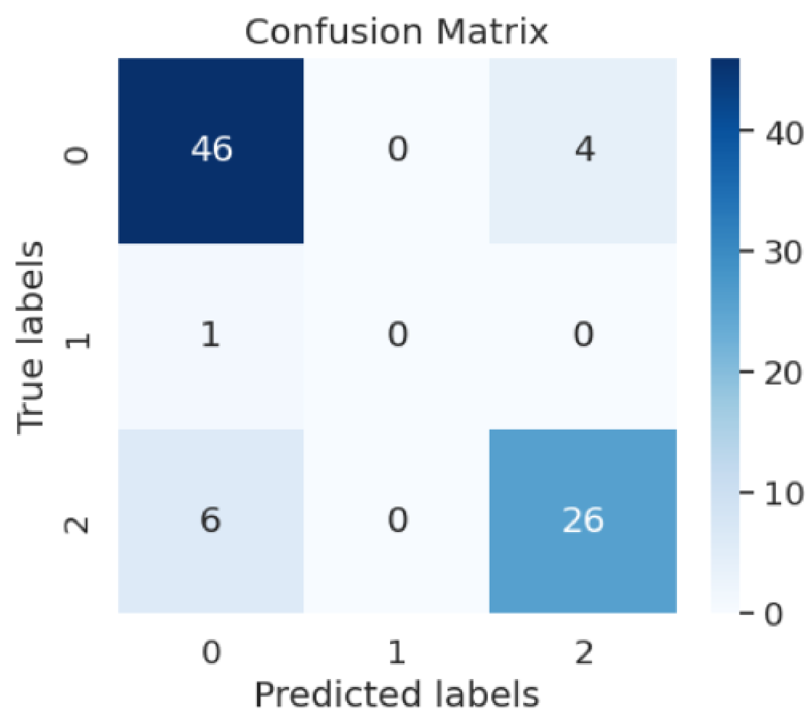

Figure 4.4: Confusion Matrix RF

SVM:

Achieved the accuracy of 60%, Precision score of 53%, Recall score of 52% and F1-score of 45%. Below at table 4.6 I have performance evaluation of SVM:

TABLE 4.5: PERFORMANCE EVALUATION(SVM)

	Precision	Recall	F1-Score	Support
C	0.62	0.94	0.75	40
CL	0.00	0.00	0.00	1
D	0.43	0.09	0.15	32
Accuracy			0.60	83
Macro avg	0.35	0.34	0.30	83
Weighted avg	0.54	0.60	0.51	83

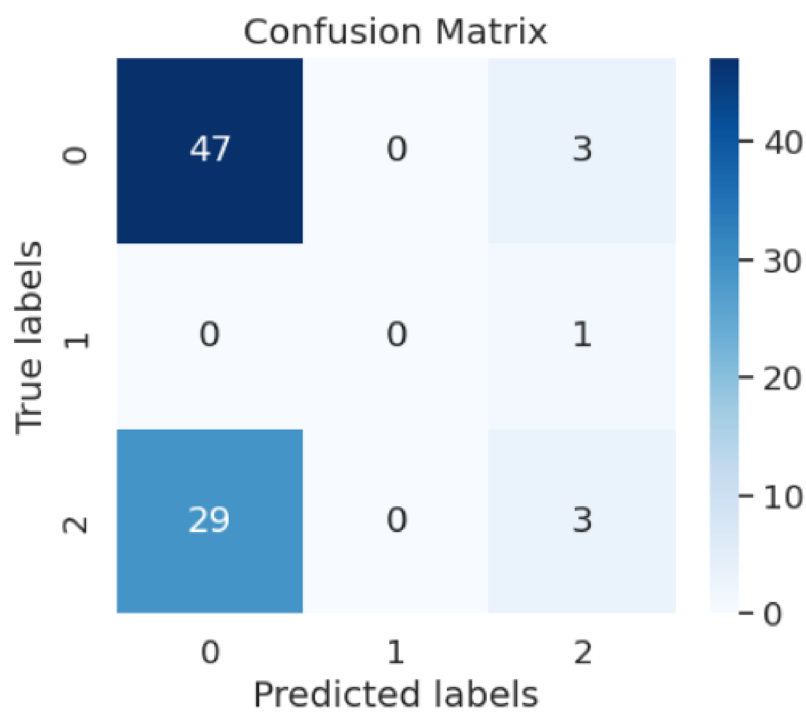

Figure 4.5: Confusion Matrix SVM

Logistic Regression:

Achieved the accuracy of 82% Precision score of 83%, Recall score of 82% and F1-score of 82%. Below at table 4.6 I have performance evaluation of LR:

TABLE 4.6: PERFORMANCE EVALUATION(LR)

	Precision	Recall	F1-Score	Support
C	0.85	0.88	0.86	50
CL	0.00	0.00	0.00	1
D	0.80	0.75	0.77	32
Accuracy			0.82	83
Macro avg	0.52	0.54	0.55	83
Weighted avg	0.82	0.82	0.82	83

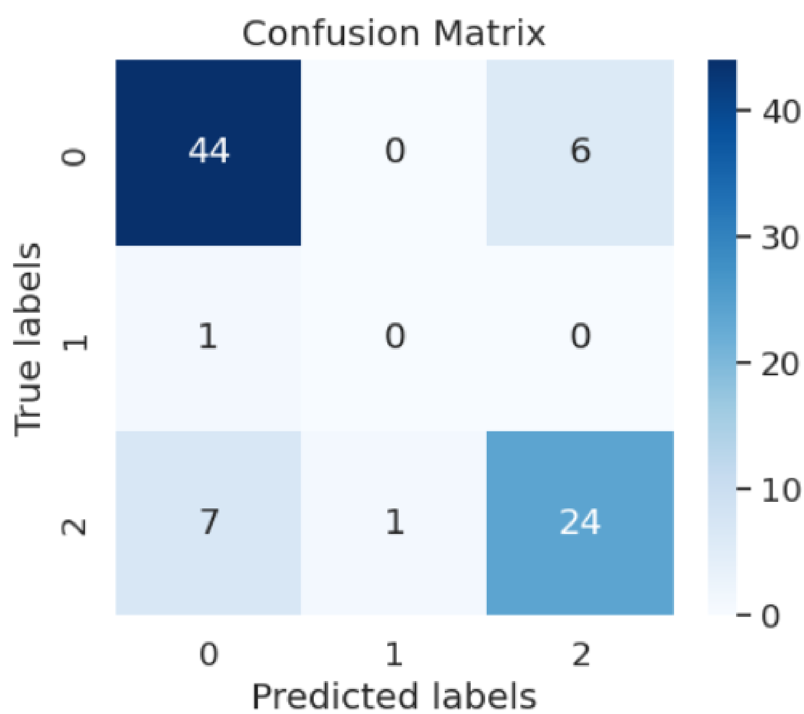

Figure 4.6: Confusion Matrix LR

4.3 Discussion

The study indicates the promise of machine learning in the prediction of liver cirrhosis. With a rich set of clinical features, the developed model showed potential in predicting the disease. The application of machine learning algorithms is particularly useful in the early diagnosis and management of liver cirrhosis. Consequently, its implementation could lead to improved patient outcomes, have positive social impacts, and reduce the economic burden on the healthcare system. Nevertheless, the model requires rigorous validation with various datasets and cohort studies to establish its robustness and generalization in response applications. Furthermore, the consideration of advanced feature selection techniques and field-specific expertise is an additional way of improving the clinical relevance and prediction accuracy of the model.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

In conclusion, predicting liver cirrhosis using machine learning techniques has the potential to revolutionize public health outcomes. The models provide accurate identification of those at risk of developing the condition, thus early interventions, and personalized medicine. The application of predictive models is beneficial to individuals and healthcare systems as it advances the ability to manage medical conditions by intervening early enough to prevent its full-blown occurrence. The models and the ensuing timely medical action can be used to change one's lifestyle, combat progression, and monitor one to avoid further deterioration. Additionally, the models contribute to open discussion about the liver organ and enable communities to take immediate actions to prevent its damage. The implementation of machine learning-based models to predict liver cirrhosis will save lives, increase survived life quality, and ease the financial burden associated with it.

5.2 Impact on Environment

Machine learning prediction of liver cirrhosis is a unique opportunity to improve patient outcomes while saving medical resources and, consequently, contributing to environmental safety. Indeed, adequate screening of high-risk patients would allow practitioners to reduce the number of diagnostic tests and invasive procedures, as well as the need for hospitalization. Thus, as a result, less medical waste, energy, and carbon footprint would be created. The same is true for the optimization of medical resource use. With machine learning models, it is possible to identify groups of patients on the basis of risk and screen or treat only high-risk patients to minimize the waste of resources. This optimization leads to a more sustainable healthcare system that conserves resources, reduces environmental footprint, and contributes to overall environmental well-being.

In conclusion, the integration of machine learning in liver cirrhosis prediction not only improves patient outcomes and healthcare efficiency but also fosters environmental sustainability by minimizing waste and resource consumption in the healthcare sector.

5.3 Sustainability Plan

Ensuring the sustainability of the Liver Cirrhosis Prediction model involves several key steps. Firstly, continual monitoring and updating of the model will be conducted to adapt to any changes in data patterns or medical knowledge. Secondly, efforts will be made to optimize computational resources to minimize energy consumption and reduce the carbon footprint associated with model training and deployment. Thirdly, collaboration with medical professionals will be maintained to ensure the model remains clinically relevant and useful. Finally, the code and documentation will be made publicly available to promote transparency, reproducibility, and further research in the field.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary of the Study

This research seeks to explore the use of machine learning models in predicting liver cirrhosis. The research is based on a dataset that contains clinical and demographic vectors. Different machine learning models are tested, and a comparison is made. The aim is to identify the best predictive model through experimenting on the test data and compare the results with the actual patient cases. The significance of this research seeks to add a perspective on attempting to predict liver cirrhosis before manifestation. Therefore, the results will provide sufficient results for predicting liver cirrhosis, making it easier to intervene early enough.

6.2 Conclusions

Since LD is a deadly disease, its treatment must be promptly provided to prevent future crises. The analysis of the model designed by ML can improve LD recognition and prevent potential long-term health effects effectively. Several ML methods are analyzed in the capacity of LD contagion predictors. Random forest classification appears as the optimal method in this case in view of the estimated indicators. The percentage of the accuracy of the models used is higher than has been in other cases, proving of the significance and reliability of the applied models. The random forest algorithm works better when cross-validation is applied to brain strokes' predictions. Several machine learning models may improve future framework models. This upgradation will improve the dependability and appearance of the framework. An approach to a machine-learning architecture enabled us to determine if an adult patient is likely to have a stroke, or not a danger sign for the general public. In such an ideal world, LD patients are treated early and live with them for the rest of their healthy lives.

6.3 Implication for Further Study

Finally, future studies can develop the model by extending the dataset with a larger selection of demographic and clinical variables. In this way, the model's power and ability to predict outcomes can be refined and become more accurate. This can also be achieved through more advanced machine learning techniques like ensemble methods, deep learning architectures, and feature engineering and selection. At a later stage, aspiring studies can track the model's performance over time through more longitudinal studies, which investigate its efficacy while linking real-world clinical settings.

REFERENCES

[1] World Health Organization, “Liver cirrhosis [Liver cirrhosis],” WHO, 2020. [Online]. Available: https://www.

[2] who.int/newsroom/fact-sheets/detail/hepatitis-c#: ∼:text=Key facts, major cause of liver cancer [Accessed: 30-Aug-2020].

[3] Sreejith, S., H. Khanna Nehemiah, and Arputharaj Kannan. "Clinical data classification using an enhanced SMOTE and chaotic evolutionary feature selection." *Computers in Biology and Medicine* 126 (2020): 103991.

[4] Stoean, Ruxandra, et al. "Evolutionary-driven support vector machines for determining the degree of liver fibrosis in chronic hepatitis C." *Artificial intelligence in medicine* 51.1 (2011): 53-65.

[5] Pellegrino, Edmund D., and H. TRISTRAM ENGELHARDT JR. "Medicine." *Philosophy, and Man’s Infirmity in Conditio Hu* (1966).

[6] Mohamed, Amal Ahmed, et al. "Hepatitis C virus: A global view." *World journal of hepatology* 7.26 (2015): 2676.

[7] Huang, Rui, et al. "Noninvasive measurements predict liver fibrosis well in hepatitis C virus patients after direct-acting antiviral therapy." *Digestive diseases and sciences* 65 (2020): 1491-1500.

[8] Cheng, Zhengjun, Yuntao Zhang, and Changhong Zhou. "QSAR Models for Phosphoramidate Prodrugs of 2′-Methylcytidine as Inhibitors of Hepatitis C Virus Based on PSO Boosting." *Chemical Biology & Drug Design* 78.6 (2011): 948-959.

[9] Singh, Lokesh, Rekh Ram Janghel, and Satya Prakash Sahu. "Classification of hepatic disease using machine learning algorithms." *Advances in Biomedical Engineering and Technology: Select Proceedings of ICBEST 2018*. Springer Singapore, 2021.

[10] Vergniol, Julien, et al. "Noninvasive tests for fibrosis and liver stiffness predict 5-year outcomes of patients with chronic hepatitis C." *Gastroenterology* 140.7 (2011): 1970-1979.

[11] Shen, Jiayi, et al. "Artificial intelligence versus clinicians in disease diagnosis: systematic review." *JMIR medical informatics* 7.3 (2019): e10010.

[12] Murawaki, Yoshikazu, et al. "Diagnostic value of serum markers of connective tissue turnover for predicting histological staging and grading in patients with chronic hepatitis C." *Journal of gastroenterology* 36 (2001): 399-406.

[13] Lackner, Carolin, et al. "Comparison and validation of simple noninvasive tests for prediction of fibrosis in chronic hepatitis C." *Hepatology* 41.6 (2005): 1376-1382.

[14] Wai, Chun-Tao, et al. "A simple noninvasive index can predict both significant fibrosis and cirrhosis in patients with chronic hepatitis C." *Hepatology* 38.2 (2003): 518-526.

[15] Halfon, Philippe, et al. "Accuracy of hyaluronic acid level for predicting liver fibrosis stages in patients with hepatitis C virus." *Comparative hepatology* 4 (2005): 1-7.

[16] "What are the stages of Liver cirrhosis?-HepatitisC.net", Healthline Media, 2018, [online] Available: https://hepatitisc.net/living/what-are-the-stages-of-hepatitis-c/.

[17] "Early detection early treatment for Liver cirrhosis", *MHealth Lab*, 2016, [online] Available: https://labblog.uofmhealth.org/rounds/early-detection-early-treatment-for-hepatitis-c.

[18] Gupta, Ekta, Meenu Bajpai, and Aashish Choudhary. "Hepatitis C virus: Screening, diagnosis, and interpretation of laboratory assays." *Asian journal of transfusion science* 8.1 (2014): 19-25.

[19] Jiang, Fei, et al. "Artificial intelligence in healthcare: past, present and future." *Stroke and vascular neurology* 2.4 (2017).

[20] Papadopoulos, Nikolaos, et al. "Liver fibrosis staging with combination of APRI and FIB-4 scoring systems in chronic hepatitis C as an alternative to transient elastography." *Annals of gastroenterology* 32.5 (2019): 498.

[21] Zhao, Yue, et al. "Novel non-invasive score to predict cirrhosis in the era of hepatitis C elimination: a population study of ex-substance users in Singapore." *Hepatobiliary & Pancreatic Diseases International* 18.2 (2019): 143-148.

[22] Kuzhippallil, Maria Alex, Carolyn Joseph, and A. Kannan. "Comparative analysis of machine learning techniques for indian liver disease patients." *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*. IEEE, 2020.

[23] Rahman, A. Sazzadur, et al. "A comparative study on liver disease prediction using supervised machine learning algorithms." *International Journal of Scientific & Technology Research* 8.11 (2019): 419-422.

[24] Arshad, Insha, et al. "Liver disease detection due to excessive alcoholism using data mining techniques." *2018 International Conference on Advances in Computing and Communication Engineering (ICACCE)*. IEEE, 2018.

[25] Ramana, Bendi Venkata, M. Surendra Prasad Babu, and N. B. Venkateswarlu. "A critical study of selected classification algorithms for liver disease diagnosis." *International Journal of Database Management Systems* 3.2 (2011): 101-114.

[26] Gregory, Onwodi. "Prediction of liver disease (biliary cirrhosis) using data mining technique." *International Journal of Emerging Technology & Research* 10.2 (2015): 37-42.

[27] Rahman, A. Sazzadur, et al. "A comparative study on liver disease prediction using supervised machine learning algorithms." *International Journal of Scientific & Technology Research* 8.11 (2019): 419-422.

[28] Jamila, Ganty, et al. "A diagnostic model for the prediction of liver cirrhosis using machine learning techniques." *Computer Science & IT Research Journal* 3.1 (2022): 36-51.

[29] Wu, Chieh-Chen, et al. "Prediction of fatty liver disease using machine learning algorithms." *Computer methods and programs in biomedicine* 170 (2019): 23-29.

[30] Singh, Jagdeep, Sachin Bagga, and Ranjodh Kaur. "Software-based prediction of liver disease with feature selection and classification techniques." *Procedia Computer Science* 167 (2020): 1970-1980.

[31] Bhupathi, Deepika, et al. "Liver disease detection using machine learning techniques." (2022).

[32] Dritsas, Elias, and Maria Trigka. "Supervised machine learning models for liver disease risk prediction." *Computers* 12.1 (2023): 19.

[33] Gogi, Vyshali J., and M. N. Vijayalakshmi. "Prognosis of liver disease: Using Machine Learning algorithms." *2018 International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE)*. IEEE, 2018.

[34] Hanif, Ishtiaqe, and Mohammad Monirujjaman Khan. "Liver Cirrhosis Prediction using Machine Learning Approaches." *2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*. IEEE, 2022.

[35] Li, Xiaoguo, et al. "Artificial intelligence in the diagnosis of cirrhosis and portal hypertension." *Journal of Medical Ultrasonics* 49.3 (2022): 371-379.

APPENDIX

ML – Machine Learning

SVM – Support Vector Machine

GB – Gradient Bosting

LR – Linear Regression

RF – Random Forest

DT – Decision Tree

PLAGIARISM REPORT

193-15-13544

ORIGINALITY REPORT

17%	14%	6%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	3%
2	www.fepbl.com Internet Source	1%
3	techscience.com Internet Source	1%
4	Ruhul Amin, Rubia Yasmin, Sabba Ruhi, Md Habibur Rahman, Md Shamim Reza. "Prediction of chronic liver disease patients using integrated projection based statistical feature extraction with machine learning algorithms", Informatics in Medicine Unlocked, 2022 Publication	1%
5	www.iieta.org Internet Source	1%
6	Submitted to Manukau Institute of Technology Student Paper	<1%
7	mro.massey.ac.nz	