ANNEXURE-1

Optimizing Liver Disease Diagnosis: A Study of Feature Selection Methods and Embedded Machine Learning Approaches

A Project Report

Submitted by:

Rahul Raj (2041018152)

Himanshu Ranjan (2041018156)

Harshit Malviya (2041013047)

Suneet Kumar (2041011132)

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHONOLOGY

IN

COMPUTER SCIENCE & INFORMATION TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY
Faculty of Engineering and Technology, Institute of Technical Education and Research
SIKSHA 'O' ANUSANDHAN (DEEMED TO BE) UNIVERSITY
Bhubaneswar, Odisha, India
(June 2024)

ANNEXURE-2

CERTIFICATE

This is to certify that the project report titled "Optimizing Liver Disease Diagnosis: A Study of Feature Selection Methods and Embedded Machine Learning Approaches" being submitted by Rahul Raj, Himanshu Ranjan, Harshit Malviya, Suneet Kumar (CSIT-D) to the Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be) University, Bhubaneswar for the partial fulfillment for the degree of Bachelor of Technology in Computer Science and Information Technology is a record of original confide work carried out by them under my supervision and guidance. The project work, in my opinion, has reached the requisite standard fulfilling the requirements for the degree of Bachelor of Technology.

The results contained in this report have not been submitted in part or full to any other University or Institute for the award of any degree or diploma.

samrudhi

Dr. Samrudhi Mohdiwale

(Name of Supervisor)

Department of Computer Science and Information Technology

Faculty of Engineering and Technology;

Institute of Technical Education and Research;

Siksha 'O' Anusandhan (Deemed to be) University

iiANNEXURE-3

ACKNOWLEDGEMENT

We are honored to present our project on **Optimizing Liver Disease Diagnosis: A Study of Feature Selection Methods and Embedded Machine Learning Approaches**. This endeavor has been a significant learning experience and an opportunity to apply our academic knowledge in a practical setting. We extend our heartfelt gratitude to several individuals and institutions who have supported and guided us throughout this journey.

First and foremost, we would like to express our sincere thanks to Institute of Technical Education and Research for providing us with the necessary resources and a conducive environment for learning and research. The institution's commitment to academic excellence has been a constant source of inspiration for us.

We are deeply grateful to our project coordinator Mr. Dilip Subuddhi, whose unwavering support and insightful feedback have been instrumental in shaping the direction and scope of this project. Their dedication to mentoring and their expertise have been invaluable.

Lastly, we wish to extend our deepest appreciation to our supervisor, Dr. Samrudhi Mohdiwale. Their guidance, encouragement, and critical insights have been pivotal in the successful completion of this project. Their expertise and patience have significantly contributed to our professional growth and the overall quality of our work.

Thank you all for your support and belief in our capabilities. We hope this project reflects the knowledge and skills imparted to us and contributes meaningfully to our field of study.

Place: ITER

Date: 29-05-2023

Signature of students

Rahul Raj

Himanshu Ranjan

Harshit Malviya

Suneet Kumar

ANNEXURE-4

DECLARATION

We declare that this written submission represents our ideas in our own words and where other's ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/fact/source in our submission. We understand that any violation of the above will cause for disciplinary action by the University and can also evoke penal action from the sources which have not been properly cited or from whom proper permission has not been taken when needed.

Signature of Students with Registration Numbers

Rahul Raj (2041018152)

Date: 29-05-2023

Himanshu Ranjan (2041018156)

Harshit Malviya (2041013047)

Suneet Kumar (2041011132)

ANNEXURE-5

REPORT APPROVAL

This project report entitled ” **Optimizing Liver Disease Diagnosis: A Study of Feature Selection Methods and Embedded Machine Learning Approaches**” by Rahul Raj, Himanshu Ranjan, Harshit Malviya, Suneet Kumar (CSIT-D) is approved for the degree of Bachelor of Technology in Computer Science & Information Technology.

Examiners

_____

_____

Abstract

Chronic liver disease (CLD) poses a significant health threat globally, with early diagnosis being crucial for effective treatment and management. This study investigates a diverse range of ensemble, boosting, and deep learning techniques for predicting CLD, emphasizing the importance of feature elimination to enhance performance and reduce computational time. Utilizing two publicly available CLD datasets, the Liver Disease Patient Dataset (LDPD) and the Indian Liver Disease Patient Dataset (ILPD), this research conducts exploratory data analysis and employs hyperparameter tuning, normalization, and upsampling for improved predictive analytics. The performance of various models, including Random Forest Classifier (RFC), Gradient Boosting (GB), AdaBoost, Multi-Layer Perceptron (MLP), and Long Short-Term Memory (LSTM), is evaluated using 10-fold cross-validation to mitigate overfitting.

The results indicate that RFC outperforms all other models with an impressive accuracy of 99.5% and a runtime approximately one-fifth that of Gradient Boosting, which has a comparable accuracy. AdaBoost achieves an accuracy of 71.1%, MLP reaches 87%, and LSTM attains 77.0%. Feature elimination significantly contributes to optimizing model performance and reducing time consumption. Among the ensemble methods, RFC emerges as the most effective technique for CLD prediction, demonstrating superior accuracy and efficiency.

	COURSE OUTCOME
CO1	Apply the acquired technical knowledge and skills to plan, analyze, design, and implement a software project or gather knowledge over the chosen field of research, and design or model solutions to proposed work.
CO2	Apply standard practices and strategies in software project development in designing and implementing a working, medium sized project as a team through simulation and/or experimental studies using modern tools and techniques.
CO3	Identify, analyze, formulate, and solve real-world problems creatively through sustainable critical investigation.
CO4	Demonstrate the ability to work as a team and communicate effectively in speech and writing through presentations and project reports.
CO5	Acquire the skills, diligence, and commitment to excellence needed to engage in lifelong learning.
CO6	Demonstrate an awareness and application of appropriate personal, societal, technical, and professional ethical standards in project development and management.

	Program Outcomes (PO) and Program Specific Outcomes (PSO)
POs	**Description**
PO1	**Engineering knowledge:** Apply the knowledge of mathematics, science, engineering fundamentals, and anengineering specialization to the solution of complex engineering problems.
PO2	**Problem analysis:** Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.
PO3	**Design/development of solutions:** Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health andsafety, and the cultural, societal, and environmental considerations.
PO4	**Conduct investigations of complex problems:** Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.
PO5	**Modern tool usage:** Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modelling to complex engineering activities with an understanding of the limitations.
PO6	**The engineer and society:** Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues, and the consequent responsibilities relevant to the professional engineering practice.
PO7	**Environment and sustainability:** Understand the impact of the professional engineering solutions in societaland environmental contexts, and demonstrate the knowledge of, and need for sustainable development.
PO8	**Ethics:** Apply ethical principles and commit to professional ethics and responsibilities and norms of theengineering practice.
PO9	**Individual and teamwork:** Function effectively as an individual, and as a member or leader in diverse teams,and in multidisciplinary settings.
PO10	**Communication:** Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and designdocumentation, make effective presentations, and give and receive clear instructions.
PO11	**Project management and finance:** Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.
PO12	**Life-long learning:** Recognize the need for and have the preparation and ability to engage in independent andlife-long learning in the broadest context of technological change.
PSO1	**Modelling and Analysis:** An ability to mathematically model and analyze the performance of ElectricalMachines, Control systems, Instrumentation systems, Power systems and Power Electronic systems.
PSO2	**Design and Development:** An ability to design the hardware and software requirements for the developmentof Electric drives Automation and Embedded systems.

LIST OF FIGURES

List of Graphs and Tables

ANNEXURE-6

Table of Contents

SDP Report

Chapter 1

1.1.1 Introduction and Background

Chronic liver disease, or CLD, is a serious health issue that affects over a million people worldwide each year[1]. About 3.5% of deaths in recent times are due to liver-related problems[1]. CLD includes various conditions like hepatocellular carcinoma (a type of liver cancer), cirrhosis (scarring of the liver), and fatty liver disease. These conditions can lead to liver failure and potentially death if not managed properly[2].

Figure 1.1: Position of the Liver in the Human Body

The liver is located in the upper right abdomen, below the diaphragm, and protected by the ribcage.Regular heavy drinking damages the liver by overworking it, leading to scarring (cirrhosis) and liver failure.

Symptoms include jaundice, swelling, and fatigue. Over time, the liver cannot function properly, posing severe health risks.[1], [3]

Misusing drugs overwhelms the liver with toxins, causing conditions like hepatitis, fibrosis, and cirrhosis. This damage results in symptoms such as nausea, loss of appetite, and liver pain. Long-term abuse can lead to severe liver failure.[3]

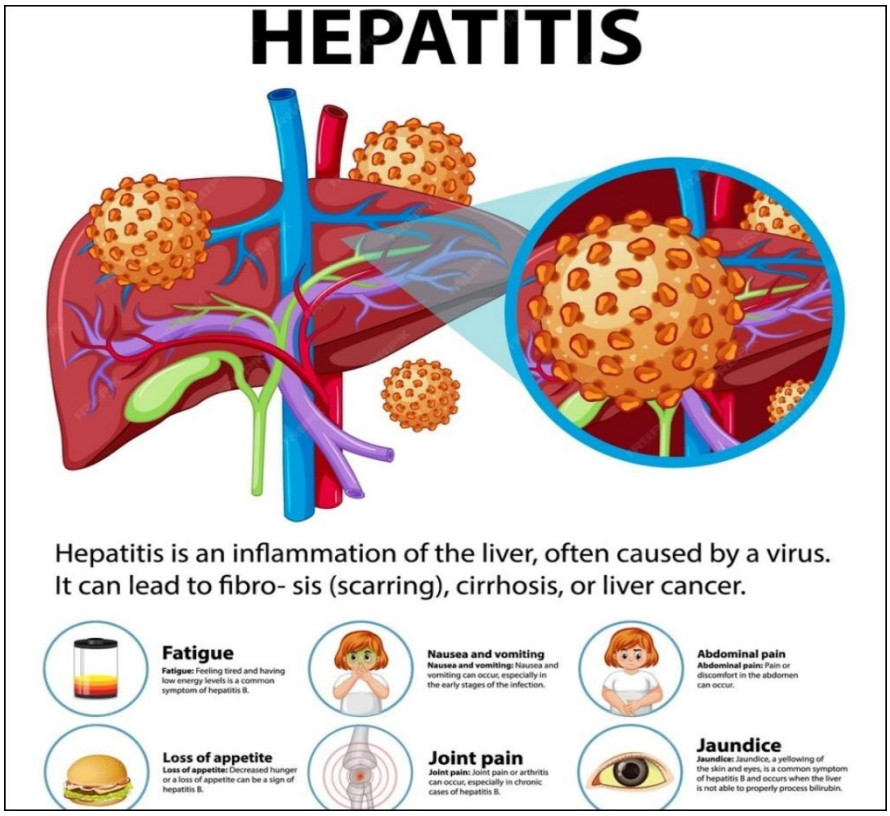


Figure 1.2: Symptoms of Hepatitis

This illustration depicts common hepatitis symptoms, including jaundice, fatigue, abdominal pain, nausea, dark urine, and loss of appetite.

Excess weight leads to fat accumulation in liver cells, causing non-alcoholic fatty liver disease (NAFLD). This can progress to severe conditions like NASH and cirrhosis, with symptoms like fatigue and abdominal pain. Weight management is crucial[1], [3], [4].

Fig 1.3 NAFLD vs Healthy liver

Type 2 diabetes causes high blood sugar and insulin resistance, leading to fat buildup in the liver (NAFLD). This can progress to NASH and cirrhosis, with symptoms including fatigue and jaundice. Effective diabetes management helps prevent liver damage.[5], [6], [7], [8]

Given the significant impact of CLD globally, it's crucial to find ways to predict and prevent it. Early detection of CLD allows for timely treatment and can improve health outcomes, potentially preventing serious complications.[7] Predicting who might develop CLD can help doctors provide personalized care and preventative measures to high-risk individuals. Chronic liver disease, or CLD, is a serious health issue that affects over a million people worldwide each year. [1], [2], [3], [9]

About 3.5% of deaths in recent times are due to liver-related problems. CLD includes various conditions like hepatocellular carcinoma (a type of liver cancer), cirrhosis (scarring of the liver), and fatty liver disease. These conditions can lead to liver failure and potentially death if not managed properly.[1], [6], [7]

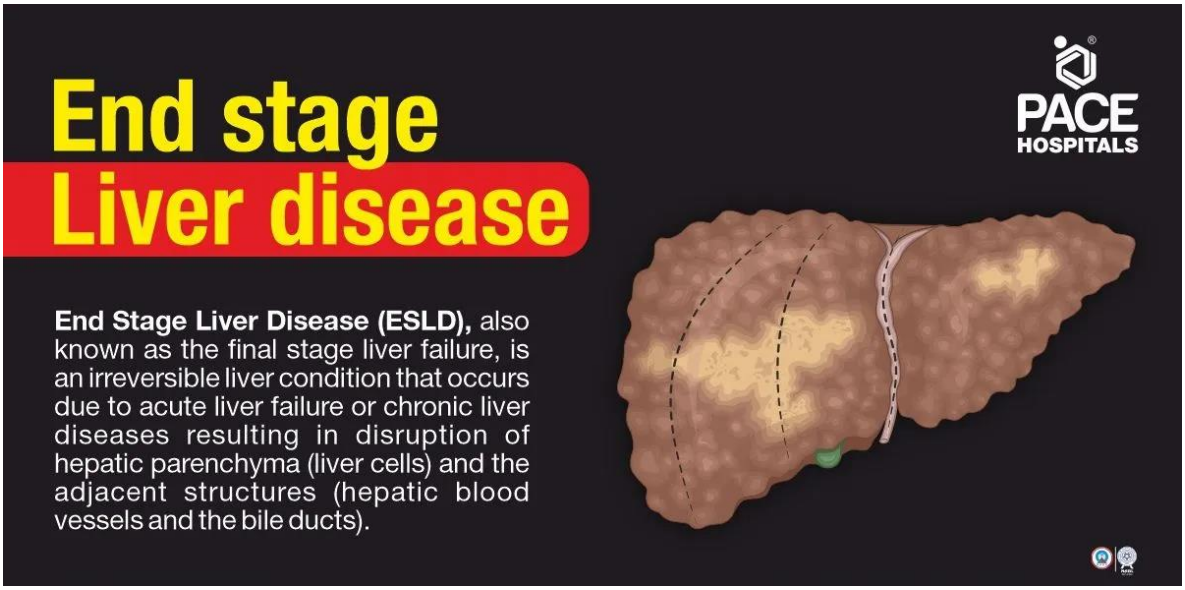


Fig 1.4 End Stage Liver Disease

Computational intelligence, particularly machine learning (ML), can play a significant role in improving CLD prediction. Machine learning involves training computers to analyze large amounts of data and identify patterns that humans might not easily see.[10], [11] This can be very useful in healthcare, where there is a wealth of data to be analyzed.

Machine learning (ML) algorithms are utilized to process extensive clinical data, including patient demographics such as age and sex, lab results, and imaging findings like liver scans. [11], [12]Through the analysis of this data, patterns and relationships associated with the development of chronic liver disease (CLD) are identified. These identified patterns are subsequently used to construct predictive models that facilitate early disease detection and risk stratification, thereby categorizing patients based on their risk

levels. This approach enhances the capability to predict and manage CLD effectively, allowing for timely interventions and personalized care plans.[1], [10], [13], [14]

The rapid growth of healthcare data has led to challenges for traditional machine learning (ML) approaches in managing large volumes of information. Consequently, advanced techniques such as ensemble learning are employed to address these challenges.[12], [14] Ensemble learning combines multiple models to enhance prediction accuracy and manageability, making it particularly effective for handling extensive datasets. By leveraging the strengths of various algorithms, ensemble learning improves performance and scalability in processing and analyzing complex healthcare data, thereby offering a robust solution to the limitations of conventional ML methods.[2], [7], [9], [15]

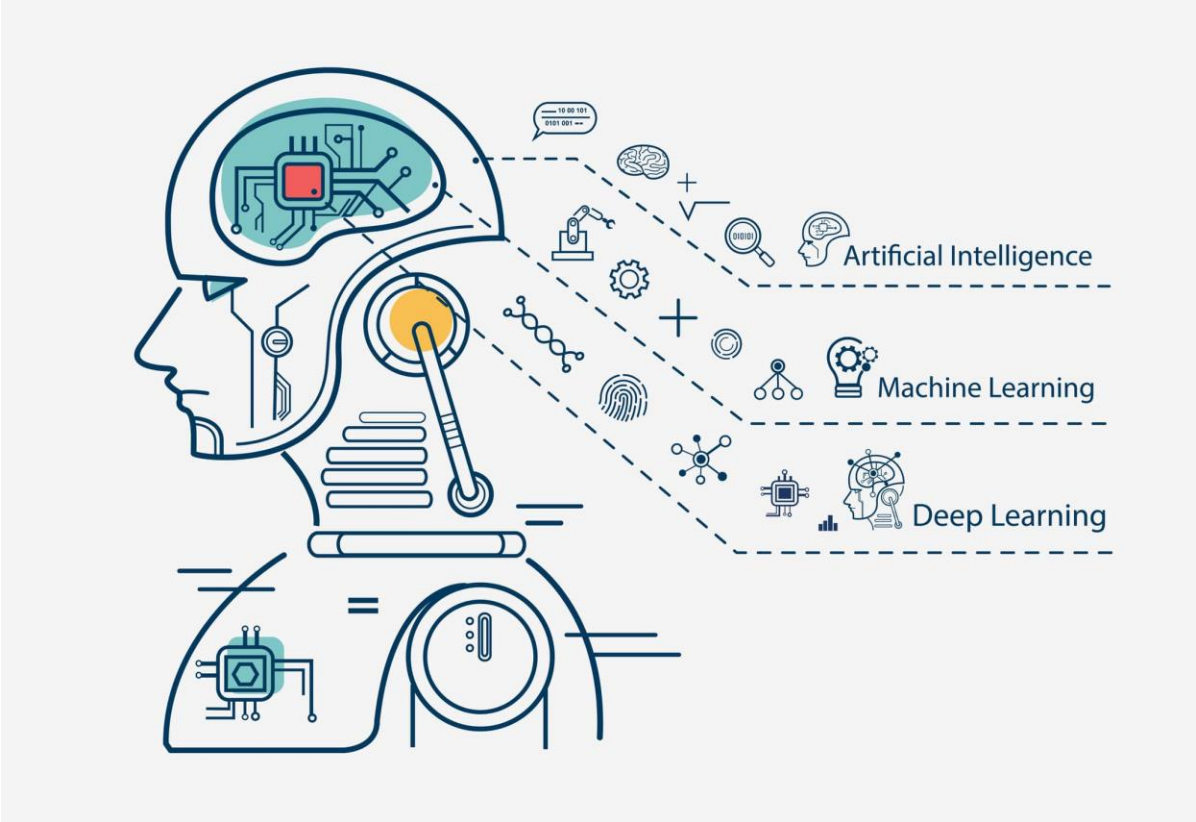


Fig 1.5 Machine Learning Design

Ensemble learning is recognized as a powerful machine learning technique in which multiple models are combined to create a single, more robust model. This method enhances the accuracy and generalizability of predictions compared to the use of individual models.[2], [7], [9], [15] By integrating the strengths of various algorithms, ensemble learning mitigates the weaknesses of individual models, leading to improved performance. The collaborative approach of ensemble learning ensures that the final model benefits from the diverse perspectives and capabilities of the combined models, resulting in more reliable and accurate predictive outcomes. [2], [7]

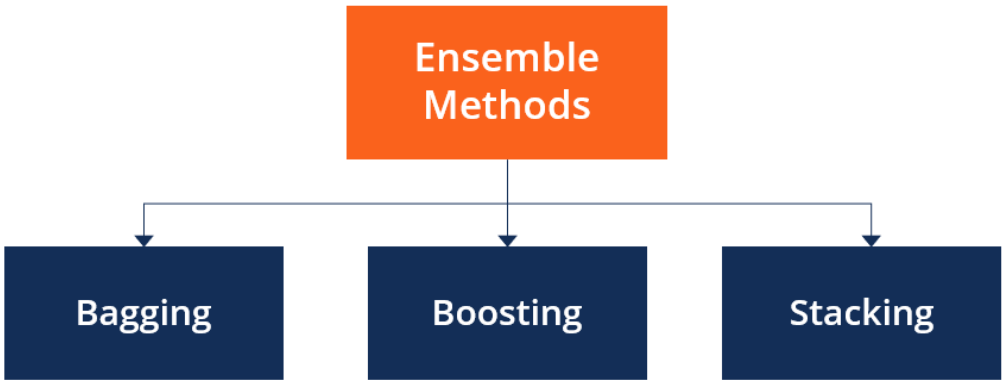


Fig 1.6 Types of Ensemble learning

Several strategies are employed in ensemble learning to enhance model performance and accuracy. Bagging involves building multiple versions of a predictor and aggregating their results by averaging or voting, which helps reduce variance and improve stability. Boosting combines several weak learners sequentially to create a strong model, with each model correcting the errors of its predecessor, thereby enhancing overall accuracy. [2], [15]Voting combines the predictions of several models by voting, where the most common prediction is selected, thus improving robustness and accuracy. Stacking involves combining multiple models by training a new model to make the final prediction based on the predictions of other models, optimizing performance through layered learning.[9]

Motivations

The motivation for this thesis stems from the critical need to improve early detection and prediction of liver disease, a condition that affects millions of people worldwide. Early and accurate diagnosis of liver disease can significantly enhance patient outcomes by enabling timely medical intervention. However, current diagnostic methods often face challenges due to the complexity and variability of liver disease symptoms, which can lead to delayed or missed diagnoses.

Advancements in machine learning present a promising solution to these challenges. By leveraging data-driven techniques, it is possible to develop predictive models that can identify patterns and indicators of liver disease more effectively than traditional methods. This thesis aims to harness the power of machine learning to create a robust and reliable prediction model for liver disease, addressing the limitations of existing diagnostic approaches.

Furthermore, the availability of large datasets and sophisticated algorithms provides an unprecedented opportunity to explore innovative approaches for disease prediction. This thesis is motivated by the potential to contribute to the medical field by improving diagnostic accuracy, reducing healthcare costs, and ultimately saving lives through early detection of liver disease. The implementation of advanced techniques like SMOTE for handling data imbalance and a thorough evaluation of different machine learning models are key components driving this research.

1.1.2 Literature review / Discussion on existing or similar Product / Process / Algorithm / Software

In recent research on liver disorder (LD) detection and prediction, various machine learning (ML) algorithms and models have been employed and compared. Shaban et al. [6] proposed an LD detection model utilizing the improved binary butterfly optimization algorithm (IB2OA), demonstrating superior performance over SVM, kNN, NB, DT, and RF with an accuracy of 98.80%, precision of 87.09%, recall of 80.10%, and an F1-score of 83.02%. Velu et al. applied Naive Bayes (NB) and the C4.5 decision tree (DT) on the ILPD dataset, with the C4.5 DT achieving a higher accuracy of 98.40%. Pasha et al. [5] developed an LD prediction model and compared its prediction accuracy with other ML algorithms such as RF, LR, and SVM.

Kalaiselvi et al. evaluated ML algorithms like kNN, DT, and ANFIS for LD prediction, with ANFIS outperforming the others on the ILPD dataset. Thirunavukkarasu et al. [8] utilized LR, kNN, and SVM for LD prediction, finding that LR and kNN achieved similar accuracy and outperformed SVM in sensitivity and specificity on the ILPD dataset. Mutlu et al. [16] developed a CNN-based model for LD detection, testing it on the BUPA and ILPD datasets, and achieving accuracies of 75.55% and 72%, respectively, while comparing its performance with NB, SVM, kNN, and LR.

Afrin et al. [17] employed ensemble learning for chronic liver disease (CLD) prediction, utilizing algorithms such as LR, DT, RF, AB, kNN, LDA, GB, and SVM. It was found that LR achieved 77.14% accuracy with all attributes, while DT achieved 94.29% accuracy, 92% precision, 99% sensitivity, and a 96% F1-score with LASSO-selected features. Amin et al. [3] developed a hybrid feature extraction method for CLD prediction and applied dimensionality reduction techniques (PCA, FA, LDA) on the ILPD dataset. The RF algorithm achieved the highest accuracy (88.1%), precision (85.33%), recall (92.3%), and an F1-score of 88.68% using 10-fold cross-validation.

Quadir et al. [2] developed an ensemble learning model with enhanced preprocessing techniques for CLD classification, employing methods such as imputation, balancing, scaling, and selection. The Extra Trees

classifier achieved the highest testing accuracy of 91.82%. Varun Vats et al. [10] compared three ML algorithms (Affinity Propagation, K-means, DBSCAN) for forecasting accuracy and computational complexity in liver disorder datasets, using the Silhouette coefficient to evaluate their comparative efficiency.

Vyshali J Gogi et al. [14] emphasized the need for advanced analytic tools to extract valuable knowledge from extensive healthcare data, focusing on liver disease. Classification algorithms such as Decision Tree, Linear Discriminant, SVM Fine Gaussian, and Logistic Regression were applied to a liver dataset using lab-based patient metrics. L. Alice Auxilia et al. [11] highlighted the global interest in using medical datasets and ML algorithms for disease prediction, employing grouping schemes and standard Indian liver illness patient records to support research in liver disease prediction.

Pushpendra Kumar et al. [13] discussed the challenges doctors face in predicting liver disorder outcomes and the class imbalance issue in liver disorder datasets. The Fuzzy-ANWKNN algorithm was introduced to improve the prediction accuracy of liver disorders. Sanjay Kumar et al. [12] analyzed various classification approaches to predict liver disorders using real-time patient datasets, evaluating the performance of five classification algorithms by measuring precision, recall, and accuracy. Lastly, Mafazalyaqeen Hassoon et al. [18] developed a new approach for rapid and accurate liver disorder diagnosis, aiding medical professionals and patients by implementing rule optimization using the Boosted C5.0 classifier and Genetic Algorithm to enhance prediction accuracy and reduce diagnosis time.

1.1.3 Problem definition

Chronic liver disease (CLD) poses a significant global health challenge, impacting millions of individuals annually and contributing to a substantial percentage of liver-related deaths. Effective prediction and early detection of CLD are critical to mitigating its severe consequences, yet traditional diagnostic methods often fall short in handling the complexity and volume of healthcare data required for accurate prognosis. [1], [3]Consequently, advanced machine learning (ML) techniques, particularly ensemble learning, are needed to enhance predictive accuracy and reliability. Addressing this need involves exploring diverse ML strategies to develop robust models capable of processing vast datasets and identifying early indicators of CLD, ultimately improving patient outcomes and preventative care measures.[11], [14]

Data Quality and Availability:

In the prediction of liver disease, challenges related to data quality and availability are often encountered. Patient records may be insufficient or incomplete, leading to gaps in crucial information required for accurate predictions.[9], [12], [13] Furthermore, inconsistencies in data formats and the presence of missing values can complicate the data preprocessing stage. Limited access to comprehensive datasets for training and validation purposes also poses a significant barrier, as models require diverse and extensive data to perform robustly.

Class Imbalance:

Class imbalance is a common issue in liver disease prediction scenarios. A disproportionate number of healthy cases compared to diseased ones can lead to biased models that are skewed towards predicting the majority class. [13]This imbalance makes it difficult to detect rare liver disease conditions accurately, as there are fewer instances for the model to learn from. Addressing this imbalance is crucial for developing a model that performs well across all classes.

Feature Selection:

In liver disease prediction, the selection of relevant features from a vast array of clinical and biochemical attributes is a critical step.[17] Identifying the most significant features that contribute to the prediction model is challenging, especially when dealing with a large number of potential variables. Additionally, multicollinearity among features, where two or more features are highly correlated, can complicate the modeling process and affect the accuracy of the predictions.

Model Selection and Complexity:

Choosing the most suitable machine learning model for liver disease prediction presents its own set of challenges. [4], [9], [17]The selection process must balance the complexity and interpretability of the model. Overly complex models, while potentially more accurate, can be difficult for healthcare professionals to interpret and trust. [9]Conversely, simpler models may not capture the intricate patterns within the data, leading to suboptimal predictive performance. Therefore, a careful selection process is essential to ensure both accuracy and usability.

Overfitting and Underfitting:

The issues of overfitting and underfitting are prevalent in the development of predictive models for liver disease.[18] Overfitting occurs when a model captures noise in the training data instead of the underlying pattern, resulting in poor generalization to new, unseen data. On the other hand, underfitting happens when

a model is too simplistic and fails to capture the complexity of the disease, leading to inaccurate predictions. Balancing these two aspects is critical to developing a robust model.

Generalization:

Ensuring that the predictive model generalizes well to unseen data and different patient populations is a significant concern in liver disease prediction. [5], [8]Variability in liver disease presentations across different demographics and populations can affect the model's performance. A model that performs well on one dataset may not necessarily perform well on another due to differences in patient characteristics and disease manifestations. Therefore, efforts must be made to validate and test the model across diverse datasets to ensure its robustness and reliability.

Objectives of our work

The objectives of this research encompass several key steps, beginning with exploratory data analysis (EDA). [12]This involves calculating statistical measures such as mean, median, and mode to understand the central tendency of the data, as well as studying the deviation to assess the spread and variability. Efficient techniques for handling missing values will be employed to ensure the dataset is clean and reliable for further analysis.

To enhance the overall classification results, optimal feature selection will be performed. [8]This process will involve both filter and wrapper methods to evaluate the relevance of different features. Recursive Feature Elimination (RFE) will be used to systematically remove unnecessary features, thus refining the feature set to those most impactful for the predictive model.

Exploration of deep learning models will also be a significant objective. Advanced boosting algorithms[1] and Random Forest Tree algorithms will be investigated for their ability to improve classification accuracy. Additionally, complex models such as Multi-Layer Perceptron (MLP) and Long Short-Term Memory (LSTM) networks will be explored to efficiently capture sequential patterns within the data, which can be crucial for accurate feature classification and prediction in liver disease scenarios.

1.1.4 Work Plan

The thesis on liver disease prediction will follow a structured approach, encompassing various essential phases from data collection to final presentation. The first week will focus on gathering relevant and high-quality datasets necessary for the prediction of liver diseases. Appropriate datasets such as LDPD and ILPD

will be identified and downloaded from reliable sources like the UCI Machine Learning Repository. The integrity and completeness of these datasets will be verified to ensure they are fit for use.

In the second week, data preprocessing will be undertaken to prepare the datasets for model building. This will involve handling missing values and outliers, normalizing or standardizing the data, encoding categorical variables, and splitting the dataset into training and testing sets.

The model building and evaluation phase will span three weeks. During the third week, initial models will be constructed using algorithms such as Boosting Algorithms, Random Forest (RF), and Neural Network Models. In the fourth week, these models will be trained on the training dataset, and their parameters will be optimized using techniques like cross-validation and grid search. The fifth week will be dedicated to evaluating the models using metrics such as accuracy, precision, recall, and F1-score, followed by a comparison of the performance of different models to select the best one.

The report and presentation phase will be ongoing throughout weeks six to eight. In the sixth week, the thesis findings will be documented, and a draft of the thesis report, including sections on introduction, methodology, results, and discussion, will be prepared. Visualizations and graphs will be created to represent data findings and model performance. The seventh week will be used to refine the report based on feedback and to develop a PowerPoint presentation summarizing the key points of the thesis. The eighth week will focus on practicing the presentation to ensure clarity and coherence, and finalizing the report and presentation materials.

This structured approach ensures that each phase of the liver disease prediction thesis is systematically covered, leading to comprehensive and well-documented findings. The summary timeline for the thesis is as follows: Week 1 for dataset collection, Week 2 for data preprocessing, Weeks 3-5 for model building and evaluation, and Weeks 6-8 for report and presentation (in progress).

Chapter 2

Design of Product/Process/Algorithm/Software

2.1 Alternative ideas

In the design of the algorithm, a comprehensive data preprocessing pipeline was implemented, encompassing feature elimination, upsampling, and data normalization. These steps were crucial for managing the dataset's complexities and ensuring effective learning by the model. K-fold cross-validation was utilized to robustly evaluate the model's performance and mitigate overfitting, resulting in more reliable and generalizable outcomes. This combination of techniques facilitated thorough preparation of the dataset, significantly enhancing the model's predictive capabilities.

Experimentation with the model without upsampling the data yielded notably less satisfactory results. The inherent class imbalance in the dataset led to biased predictions, predominantly favoring the majority class. This imbalance caused a significant reduction in recall for the minority class, resulting in many cases of the liver disorder not being correctly identified. In a healthcare context, such a lack of sensitivity in identifying the disorder is unacceptable due to the serious implications of false negatives for patient health.

Similarly, omitting feature elimination from the preprocessing stage resulted in reduced model performance. The inclusion of irrelevant or redundant features introduced noise into the model, impairing its ability to discern important patterns within the data. This noise not only diminished the model's accuracy but also increased the complexity and computational cost of the training process. The presence of extraneous features made the model more prone to overfitting, leading to poor generalization on unseen data.

Lastly, the absence of data normalization proved detrimental to the model's effectiveness. Without normalization, features with larger numerical ranges dominated those with smaller ranges, skewing the model's learning process. This imbalance resulted in suboptimal weights being assigned during training, reducing the model's overall accuracy and stability. The lack of normalization hindered the model's ability to converge to an optimal solution, leading to slower training times and less reliable predictions. Conversely, the inclusion of normalization ensured that all features contributed equally to the learning process, thereby enhancing the model's performance and robustness.

2.2 Design / Comparison criteria

For the design and comparison phase, the focus will be primarily on the selection of models and the elimination of less relevant features. Various machine learning models will be evaluated to determine which one performs best for liver disease prediction. This will involve systematically eliminating features that do not significantly contribute to the model's accuracy and efficiency. Through careful comparison, the most effective combination of features and models will be identified, ensuring that the final model is both robust and reliable. This approach will facilitate the selection of a model that not only predicts liver disease accurately but also operates efficiently with the most relevant data inputs.

2.3 Evaluation for selection of best idea

The evaluation for selecting the best model idea will primarily focus on achieving optimized results through feature elimination. By systematically removing features that do not significantly contribute to the model's predictive power, the model's performance can be enhanced. This process helps in refining the model, ensuring it utilizes only the most relevant data inputs, thereby increasing its accuracy and reliability. The ultimate goal is to develop a model that delivers precise predictions for liver disease, leveraging the most critical features to maximize its effectiveness.

In addition to optimizing results, feature elimination plays a crucial role in reducing the model's runtime. By eliminating irrelevant or redundant features, the computational load on the model is significantly decreased, leading to faster processing times. This efficiency gain is particularly important in real-world applications where quick and accurate predictions are essential. The time saved during model runtime not only improves the user experience but also allows for more rapid iterations and adjustments to the model, further enhancing its overall performance and applicability.

2.4 Detailed design

Flow Chart

The following flowchart illustrates the comprehensive data preprocessing and model evaluation pipeline designed for liver disease prediction. This process includes critical steps such as feature elimination, upsampling, and data normalization, followed by the application of k-fold cross-validation to ensure robust

performance assessment and to mitigate overfitting. Each stage is essential for preparing the dataset and enhancing the predictive accuracy of the model.

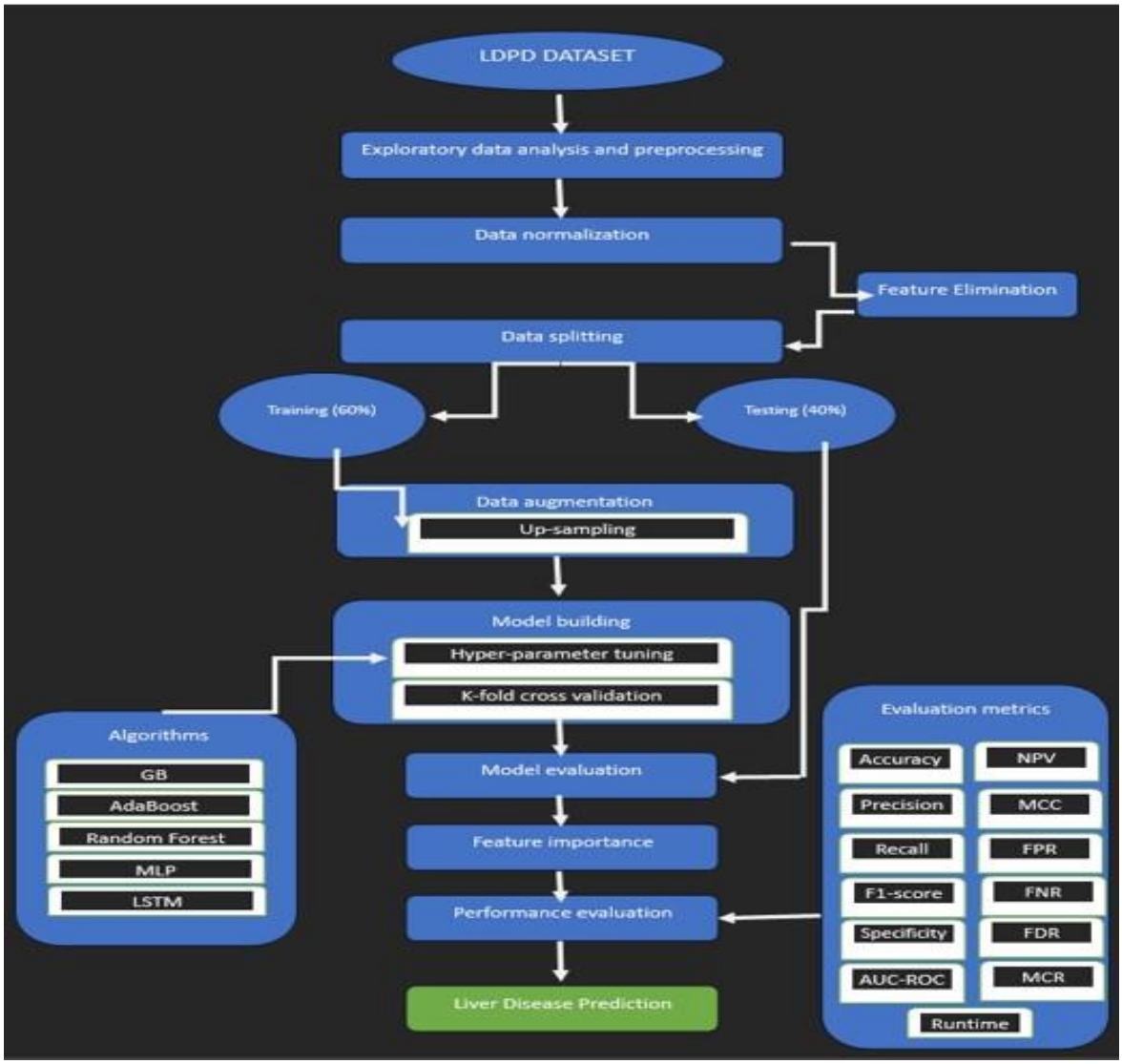


Fig 2.1 Flowchart

The datasets used in this thesis are critical for developing an accurate and reliable model for liver disease prediction. These datasets, sourced from reliable repositories, contain comprehensive data necessary for training and evaluating the machine learning models.

Dataset Description

This data set contains 10 variables that are age, gender, total Bilirubin, direct Bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT and Alkphos. Detailed descriptions and access to the datasets can be found at https://www.kaggle.com/datasets/abhi8923shriv/liver-disease-patient-dataset. [1]

Attribute Information:

1. Age of the patient
2. Gender of the patient
3. TB- Total Bilirubin
4. DB- Direct Bilirubin
5. Alkphos- Alkaline Phosphotase
6. Sgpt- Alamine Aminotransferase
7. Sgot- Aspartate Aminotransferase
8. TP- Total Protiens
9. ALB- Albumin
10. A/G Ratio- Albumin and Globulin Ratio

	Age	TB	DB	Alkphos	Sgpt	Sgot	TP	ALB	A/G Ratio	Result
count	30689.000000	30043.000000	30130.000000	29895.000000	30153.000000	30229.000000	30228.000000	30197.000000	30132.000000	30691.000000
mean	44.107205	3.370319	1.528042	289.075364	81.488641	111.469979	6.480237	3.130142	0.943467	0.714118
std	15.981043	6.255522	2.869592	238.537589	182.158850	280.851078	1.081980	0.792281	0.323164	0.451841
min	4.000000	0.400000	0.100000	63.000000	10.000000	10.000000	2.700000	0.900000	0.300000	0.000000
25%	32.000000	0.800000	0.200000	175.000000	23.000000	26.000000	5.800000	2.600000	0.700000	0.000000
50%	45.000000	1.000000	0.300000	209.000000	35.000000	42.000000	6.600000	3.100000	0.900000	1.000000
75%	55.000000	2.700000	1.300000	298.000000	62.000000	88.000000	7.200000	3.800000	1.100000	1.000000
max	90.000000	75.000000	19.700000	2110.000000	2000.000000	4929.000000	9.600000	5.500000	2.800000	1.000000

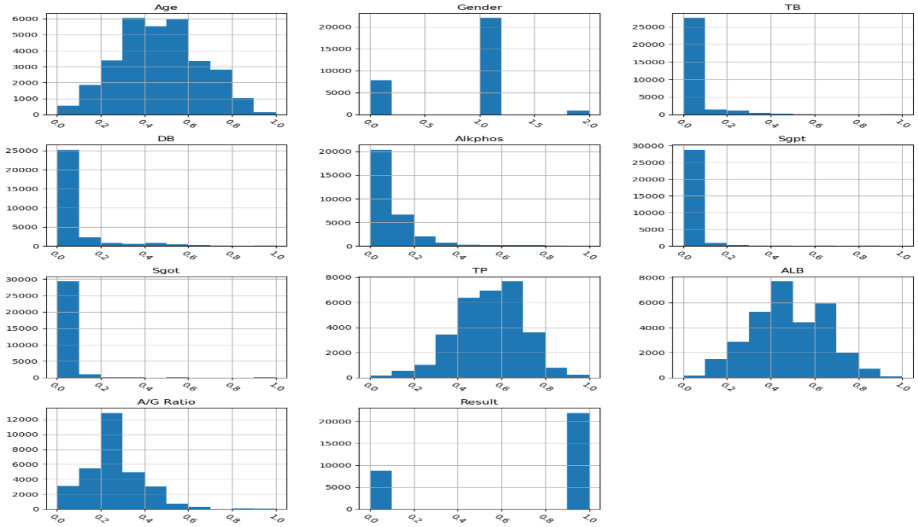


Fig 2.2 Data Description

- **Data Preprocessing-**

In the data preprocessing stage, several critical steps were undertaken to ensure the dataset was primed for effective modeling. Initially, any null or missing data points were addressed using interpolation techniques, which allowed for the seamless filling of gaps without introducing significant bias. [2]This method ensured the continuity and completeness of the dataset. Additionally, categorical variables were transformed into numerical format through label encoding, facilitating their use in machine learning algorithms. To enhance data quality further, an outlier detection process was employed to identify and potentially remove or adjust anomalous data points that could skew the results. [7]These preprocessing steps were essential in creating a clean, consistent, and reliable dataset, which is fundamental for building robust and accurate predictive models. Label encoding the features in the dataset plays a significant role in enhancing model performance, particularly for algorithms that are sensitive to the numerical representation of categorical data. Label encoding converts categorical variables into numerical values, allowing machine learning algorithms to process and interpret the data effectively.[9]

For instance, categorical features such as 'gender' or “disease result”, which are represented by strings, are transformed into numerical labels. This transformation helps the model to understand the inherent relationships and differences between the categories, which might not be possible if the data remains in its original string format. [5]Label encoding ensures that these categorical features are meaningfully incorporated into the model training process.

Moreover, some machine learning algorithms, especially tree-based methods like Random Forest and Gradient Boosting, can handle numerical data more efficiently than categorical data. By converting categorical features into numerical labels, the algorithms can perform calculations and make splits based on the numerical values, thereby improving the model's ability to identify patterns and make accurate predictions.

Overall, label encoding simplifies the dataset and enhances the model's performance by ensuring that all features are in a numerical format that the algorithms can process effectively, leading to better learning and prediction outcomes.

```
[9]:
df.isnull().sum()

[9]:
Age             2
Gender          0
TB            648
DB            561
 Alkphos      796
 Sgpt         538
Sgot          462
TP            463
 ALB          494
A/G Ratio     559
Result          0
dtype: int64
```

Fig 2.3 Null Values in the data

- **Data Normalisation –**

The MinMaxScaler algorithm scales numerical features to a specified range, typically between 0 and 1, using the formula:

$$X\_scaled = X - X\_min/(X\_max - X\_min)$$

where X is the original feature value, X_min is the minimum value of the feature, and X_max is the maximum value of the feature.

In the dataset used for this thesis, there is a significant disparity in the range of values across different features. For example, the 'alkophos' (alkaline phosphatase) levels are typically in the range of 150 and above, whereas the 'ag_ratio' (albumin-globulin ratio) values are around 1 or less. Other features, such as 'tp' (total protein) and 'sgpt' (serum glutamic-pyruvic transaminase), also exhibit much lower values. This wide variation in the scales of different features can adversely affect the performance of machine learning models, as many algorithms assume that the data is normalized.

To address this issue, data normalization was applied to bring all features onto a similar scale. [11]Normalization transforms the data into a standard range, typically between 0 and 1, by subtracting the mean and dividing by the standard deviation for each feature. This process ensures that each feature contributes equally to the model, preventing features with larger ranges from disproportionately influencing the model's learning process.

By normalizing the data, the convergence of gradient descent algorithms is accelerated, and the overall stability and performance of the models are improved. Normalization helps in maintaining numerical stability during computations and ensures that the optimization process is efficient. This step is crucial for enhancing the accuracy and reliability of the liver disease prediction models developed in this thesis.

- **Feature Classification and Elimination –**

In the feature classification [15]and elimination process, we employed several advanced techniques including ReliefF, correlation analysis, mutual information, and Recursive Feature Elimination (RFE) with a Gradient Boosting algorithm. Each of these methods was instrumental in evaluating the importance of individual features and their contribution to the predictive power of the model. ReliefF effectively ranked features by their ability to distinguish between different classes, while correlation analysis identified and removed redundant features that were highly correlated with each other. Mutual information measured the dependency between each feature and the target variable, providing insights into the most informative features.

Despite the varied methodologies, all these techniques consistently highlighted the same subset of features as the most significant. This convergence reinforced the reliability of our feature selection process, ensuring that we retained the most predictive variables. RFE was particularly useful as it iteratively removed the least important features based on the performance of the Gradient Boosting algorithm, refining the feature set to enhance model accuracy.

Ultimately, five key features were selected through this rigorous process. The use of RFE was critical in this context as it combined the strengths of feature ranking and model training, providing a robust mechanism to identify the most valuable features for the final model. This careful selection of features contributed significantly to the model's ability to make accurate and reliable predictions, confirming the effectiveness of our comprehensive feature classification and elimination strategy.

Data Correlation -

	Age	Gender	TB	DB	Alkphos	Sgpt	Sgot	TP
Age	1.000000	0.021505	-0.006060	-0.003615	0.001069	-0.003875	-0.000411	-0.006999
Gender	0.021505	1.000000	0.021749	0.023711	0.012632	0.017066	0.015511	0.001724
TB	-0.006060	0.021749	1.000000	0.865872	0.194282	0.195815	0.233436	0.002126
DB	-0.003615	0.023711	0.865872	1.000000	0.224436	0.218511	0.254019	0.009129
Alkphos	0.001069	0.012632	0.194282	0.224436	1.000000	0.113158	0.144308	-0.023215
Sgpt	-0.003875	0.017066	0.195815	0.218511	0.113158	1.000000	0.786264	-0.041855
Sgot	-0.000411	0.015511	0.233436	0.254019	0.144308	0.786264	1.000000	-0.027571
TP	-0.006999	0.001724	0.002126	0.009129	-0.023215	-0.041855	-0.027571	1.000000
ALB	-0.015332	0.001666	-0.216879	-0.224929	-0.155356	-0.024258	-0.079011	0.767084
A/G Ratio	-0.020270	0.000702	-0.203838	-0.200011	-0.224276	0.000638	-0.059871	0.223465

Fig 2.4 Data Correlation

```
Pairs of features with correlation > 0.5:
('TB', 'DB')
('Sgpt', 'Sgot')
('TP', 'ALB')
('ALB', 'A/G Ratio')
```

Fig 2.5 Pairs of Features with correlation

After final elimination using RFE, the selected features were Total Bilirubin (TB), Direct Bilirubin (DB), Alkaline Phosphatase (Alkphos), Serum Glutamate Pyruvate Transaminase (Sgpt), and Serum Glutamic-Oxaloacetic Transaminase (Sgot).

• Data Splitting:

In this thesis, the dataset was split into training and testing sets to facilitate the model building and evaluation process. A 60-40 ratio was used, with 60% of the data allocated for training the model and the remaining 40% reserved for testing its performance. [1]This approach ensured that a substantial portion of the data was utilized to develop the model, while a significant portion was held back to objectively assess its accuracy and generalization capabilities on unseen data.

Training data - 18,414 rows

Testing data - 12,277 rows

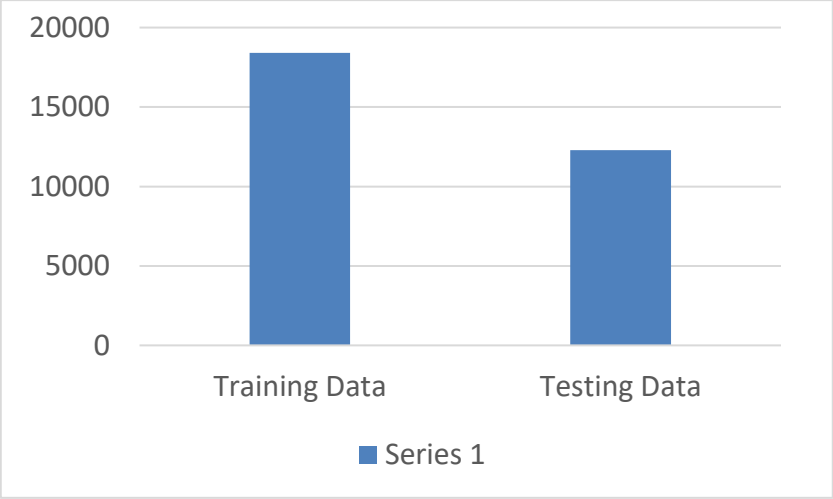


Fig 2.6 Training and Testing Data division

• Handling Data Imbalance:

To address data imbalance in the liver disease prediction dataset, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. Data imbalance, where certain classes in the dataset are underrepresented, can lead to biased and inaccurate model predictions. SMOTE tackled this issue by generating synthetic samples for the minority class, thereby balancing the dataset. This method created new instances based on the feature space similarities between existing minority class samples. By applying SMOTE, the model was trained on a more balanced dataset, improving its ability to accurately predict liver disease across all classes.

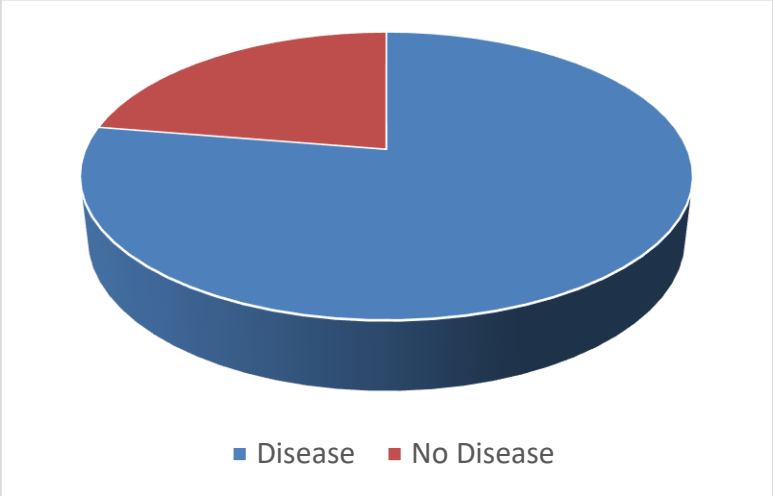


Fig 2.6 Disease and Non - Disease chart division

Initially, the dataset comprised 21,917 instances of data with disease and 8,774 instances of data without disease. To prepare the data for model building and evaluation, it was first split into training and testing sets using a 60-40 ratio. This division ensured that 60% of the data was used for training the model, while the remaining 40% was reserved for testing its performance. After splitting the data, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training set to address the class imbalance. SMOTE generated synthetic samples for the minority class, creating a more balanced dataset and enhancing the model's ability to accurately predict liver disease across all classes.

• Model Building:

In the study, we selected a diverse set of machine learning models, including Gradient Boosting (GB), AdaBoost, Random Forest Classifier, Multi-Layer Perceptron (MLP), and Long Short-Term Memory (LSTM), to ensure comprehensive coverage of different modeling approaches and their respective strengths.

Gradient Boosting (GB)[1] was chosen for its ability to sequentially build a strong learner by combining multiple weak learners, effectively capturing complex patterns in the data. Its robustness to overfitting and high predictive performance make it an excellent choice for classification tasks.

AdaBoost [1]was included as it enhances the performance of weak classifiers by focusing on misclassified instances in each iteration. This adaptive boosting method is valuable for improving model accuracy and handling diverse data distributions.

Random Forest Classifier [15] was selected for its ensemble nature, which combines multiple decision trees to improve generalization and reduce overfitting. Its capability to handle large datasets with higher dimensionality and its inherent feature importance ranking are crucial for our analysis.

Multi-Layer Perceptron (MLP) represents a class of neural networks [16] that can model non-linear relationships by learning through backpropagation. Its flexibility in architecture allows it to capture intricate patterns and interactions within the data.

Long Short-Term Memory (LSTM) networks were chosen for their proficiency in handling sequential data and long-term dependencies. Although more commonly used in time series analysis, LSTMs can uncover temporal relationships and trends that might be present in the dataset.

By employing this diverse set of models, we aimed to cover a wide spectrum of machine learning paradigms, including boosting techniques, ensemble methods, neural networks, and recurrent architectures. This comprehensive approach ensures that we leverage the unique strengths of each model type, thereby enhancing the robustness and accuracy of our predictive analysis.

- **Hyperparameter Tuning and K-fold Cross Validation** -

In the study, we employed k-fold cross-validation with k set to 10, combined with hyperparameter tuning, to enhance the accuracy and robustness of the machine learning models. This approach involved dividing the dataset into 10 equal parts, or folds. Each model was trained on 9 folds and tested on the remaining fold, repeating this process 10 times, ensuring that each fold was used as a test set exactly once.

This method provided several key benefits:

1. **Improved Accuracy and Reliability**: By averaging the performance metrics across all 10 folds, we obtained a more reliable estimate of the model's accuracy. This helped mitigate the variability that can arise from a single train-test split, leading to a more robust assessment of the model's performance.

2. Reduced Overfitting: The iterative nature of k-fold cross-validation allowed us to train and validate the models on different subsets of the data, reducing the risk of overfitting. This comprehensive training approach ensured that the models generalized better to unseen data, thereby enhancing their predictive accuracy.

By utilizing 10-fold cross-validation with hyperparameter tuning, we were able to systematically evaluate and fine-tune the models, resulting in improved performance and greater confidence in the accuracy of the predictive analyses. This rigorous approach was crucial in developing robust machine learning models capable of delivering reliable and accurate predictions.

Chapter 3

Results on Basis of Model Category

The results of the liver disease prediction models illustrate the performance of various machine learning algorithms, which can be broadly categorized into ensemble methods, boosting methods, and neural networks. Each category has distinct characteristics and mechanisms that influence their effectiveness on the dataset.

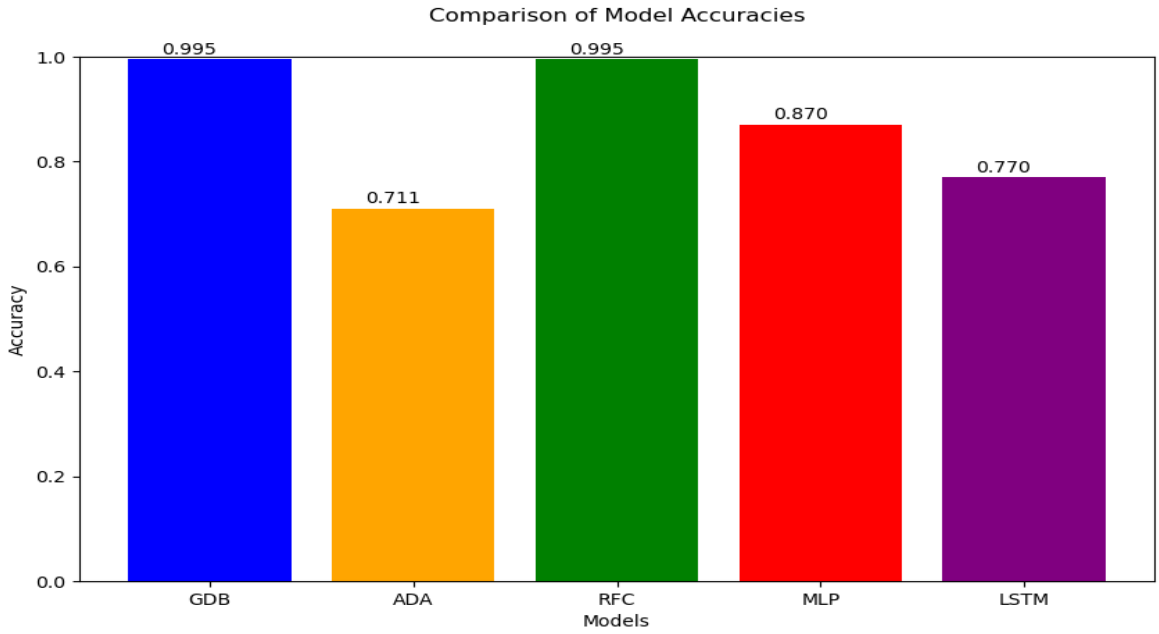


Fig 3.1 Result Comparison

Model	Accuracy	Precision	Recall	F1-Score	False Negativity Rate	False Discovery Rate	Misclassification Rate	AUC
GD Boost	**0.995**	**0.996**	**0.996**	**0.996**	**0.003**	**0.003**	**0.004**	**0.994**
Ada Boost	**0.711**	**0.935**	**0.641**	**0.761**	**0.358**	**0.064**	**0.289**	**0.764**
Random Forest Classification	**0.995**	**0.996**	**0.997**	**0.996**	**0.002**	**0.003**	**0.004**	**0.993**

Multi - layered Perceptron	**0.870**	**0.980**	**0.835**	**0.902**	**0.164**	**0.196**	**0.129**	**0.896**
Long Short Term Memory	**0.763**	**0.949**	**0.709**	**0.811**	**0.290**	**0.050**	**0.236**	**0.806**

Result Evaluation Table

Ensemble Methods:

Ensemble methods combine the predictions of multiple base models to improve overall performance. These methods are known for their ability to reduce variance and avoid overfitting, leading to more robust and accurate predictions.

1. Random Forest (RFC):

```
Total Runtime: 316.43401622772217 seconds
Best Hyperparameters (Random Forest): {'max_depth': 20, 'min_samples_leaf': 1, 'min_
samples_split': 2, 'n_estimators': 200, 'verbose': 1}
Test Accuracy (Random Forest): 0.9953571719475441
Precision: 0.9961516694963215
Recall: 0.9973934723481415
F1-score: 0.9967721841553882
Specificity: 0.9901534897190849
False Negative Rate: 0.0026065276518585675
False Discovery Rate: 0.003848330503678551
Misclassification Rate: 0.004642828052455811
AUC: 0.9937734810336132
```

Fig 3.2 RFC Result

- Test Accuracy: 0.9954

Random Forest is an ensemble method that constructs a multitude of decision trees during training and outputs the mode of the classes (classification) of the individual trees. By aggregating the results of multiple trees, Random Forest enhances prediction accuracy and reduces the risk of overfitting. In this study, it achieved an impressive test accuracy of 99.54%, indicating its strong capability to handle the liver disease dataset effectively.

Boosting Methods:

Boosting methods focus on converting weak learners into strong ones by sequentially building models that correct the errors of their predecessors. These methods are particularly powerful for handling complex datasets with subtle patterns.

1. Gradient Boosting (GB):

```
Best Hyperparameters: {'learning_rate': 0.01, 'max_depth': 8, 'n_estimators': 6000,
'subsample': 0.8, 'verbose': 1}
Test Accuracy: 0.9951942657000896
Total Runtime: 3052.790319442749 seconds
Accuracy: 0.9951942657000896
Precision: 0.9969384283932419
Recall: 0.9963735267452403
F1-score: 0.996655897523097
Specificity: 0.9921807124239791
False Negative Rate: 0.003626473254759746
False Discovery Rate: 0.003061571606758136
Misclassification Rate: 0.004805734299910402
AUC: 0.9942771195846096

[21]:
```

Fig 3.3 GD Boost Result

- Test Accuracy: 0.9952

Gradient Boosting is a boosting technique that builds models sequentially, each new model attempting to correct the errors made by the previous models. It combines these models to produce a strong learner. The Gradient Boosting model achieved a high test accuracy of 99.52%, demonstrating its efficiency in capturing complex patterns in the dataset by focusing on the hardest-to-predict instances in each iteration.

2. AdaBoost:

```
Best parameters: {'learning_rate': 0.01, 'n_estimators': 6000}
Test Accuracy (AdaBoost): 0.7105155982731938
Precision (AdaBoost): 0.9352494218698382
Recall (AdaBoost): 0.6416591115140526
F1-score (AdaBoost): 0.7611238069633015
Specificity (AdaBoost): 0.8864755285259195
False Negative Rate (AdaBoost): 0.3583408884859474
False Discovery Rate (AdaBoost): 0.06475057813016187
Misclassification Rate (AdaBoost): 0.28948440172680623
AUC (AdaBoost): 0.764067320019986
Total Runtime: 710.9173624515533 seconds
```

Fig 3.4 Ada Boost Result

- Test Accuracy: 0.7105

AdaBoost, another boosting method, works by adjusting the weights of incorrectly classified instances so that subsequent models focus more on these difficult cases. While it achieved a test accuracy of 71.05%, which is lower than Gradient Boosting and Random Forest, it still highlights the importance of boosting in improving model performance, albeit with some sensitivity to noisy data and outliers.

Neural Networks:

Neural networks are a class of models inspired by the human brain, capable of learning complex patterns through layers of interconnected nodes (neurons). They are particularly effective for tasks involving high-dimensional data and intricate relationships.

3. Multi-Layer Perceptron (MLP):

```
Best Hyperparameters (MLP): {'optimizer': 'adam', 'epochs': 40, 'batch_size': 32}
Test Accuracy (MLP): 0.8700008145312372
Precision: 0.9803296119085593
Recall: 0.8359020852221215
F1-score: 0.9023733790066063
Specificity: 0.9571387199536635
False Negative Rate: 0.16409791477787852
False Discovery Rate: 0.019670388091440724
Misclassification Rate: 0.12999918546876274
AUC: 0.8965204025878926
```

Fig 3.5 MLP Result

- Test Accuracy: 0.8700

The Multi-Layer Perceptron is a type of feedforward artificial neural network consisting of multiple layers of neurons. It is capable of learning non-linear relationships in the data. The MLP achieved a test accuracy of 87.00%, demonstrating its ability to capture complex patterns, although it was slightly less effective compared to ensemble methods. MLPs require careful tuning of hyperparameters and substantial computational resources.

4. Long Short-Term Memory (LSTM):

```
Best Hyperparameters (LSTM): {'units': 128, 'optimizer': 'adam', 'epochs': 40}
Test Accuracy (LSTM): 0.7637044880671173
Precision: 0.9490523123578468
Recall: 0.709315503173164
F1-score: 0.8118555029509047
Specificity: 0.9026933101650738
False Negative Rate: 0.2906844968268359
False Discovery Rate: 0.05094768764215315
Misclassification Rate: 0.23629551193288262
AUC: 0.806004406669119
```

Fig 3.6 LSTM Result

- Test Accuracy: 0.7630

LSTM networks, a variant of recurrent neural networks (RNNs), are designed to handle sequential data and long-term dependencies. In this study, the LSTM achieved a test accuracy of 77.00%. Despite their strengths in sequence prediction, LSTMs were not as well-suited for this particular dataset, which did not primarily consist of temporal data.

Ensemble methods, particularly Random Forest, provided the highest accuracy by aggregating the results of multiple decision trees, demonstrating robustness and effectiveness in handling diverse data patterns. Boosting methods, including Gradient Boosting and AdaBoost, also showed strong performance. Gradient Boosting, in particular, effectively captured complex patterns through iterative error correction, showcasing its capability to improve model accuracy progressively. Neural networks, represented by MLP and LSTM, demonstrated the ability to model complex relationships within the data. The MLP performed well, although it required significant tuning to optimize its parameters, whereas the LSTM was less suited for the non-sequential data in this study, highlighting its strength primarily in sequence prediction tasks.

Overall, ensemble and boosting methods, particularly Random Forest and Gradient Boosting, emerged as the most effective approaches for liver disease prediction in this dataset, providing the best balance of accuracy and robustness. Neural networks also showed potential but required further optimization to achieve comparable performance.

Results on basis of Features and Time Consumption

After narrowing down the models to Random Forest Classifier (RFC) and Gradient Boosting (GD Boost) based on accuracy, we prioritized optimizing runtime. The analysis demonstrated that both GD Boost and RFC showed significant reductions in runtime when the number of features was reduced from 10 to 5.

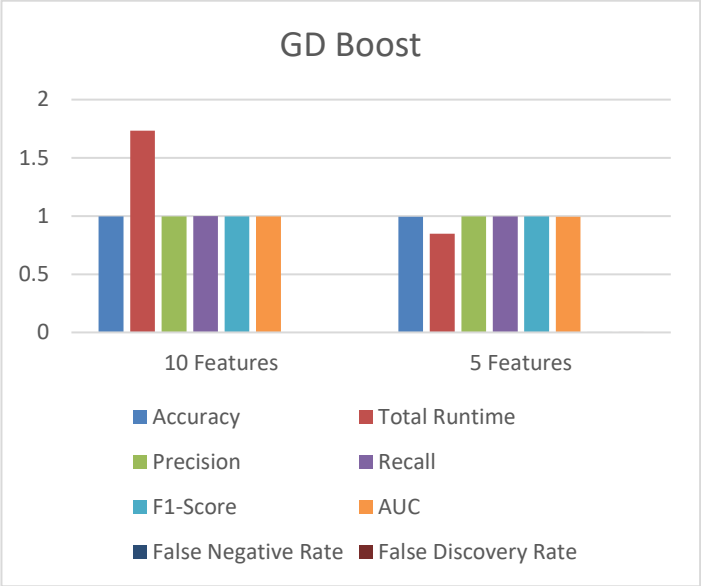


Fig 3.7 GD Boost Optimisation Comparision

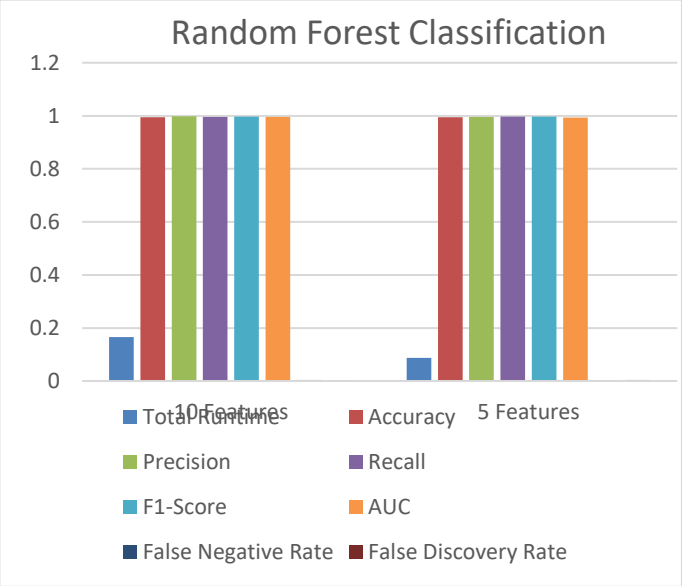


Fig 3.8 RFC Optimisation Comparison

For Gradient Boosting, the runtime decreased by approximately 51%, dropping from 6240.70 seconds to 3052.79 seconds. Similarly, for Random Forest, the runtime decreased by approximately 47%, from 601.08 seconds to 316.43 seconds. These substantial reductions in runtime highlight the effectiveness of feature elimination in making the models more efficient.

Despite reducing the number of features, the performance metrics (Accuracy, Precision, Recall, F1-Score, AUC, False Negative Rate, and False Discovery Rate) remained relatively stable for both GD Boost and Random Forest. The slight decreases observed in these metrics were marginal and considered a reasonable trade-off for the significant decrease in computation time.

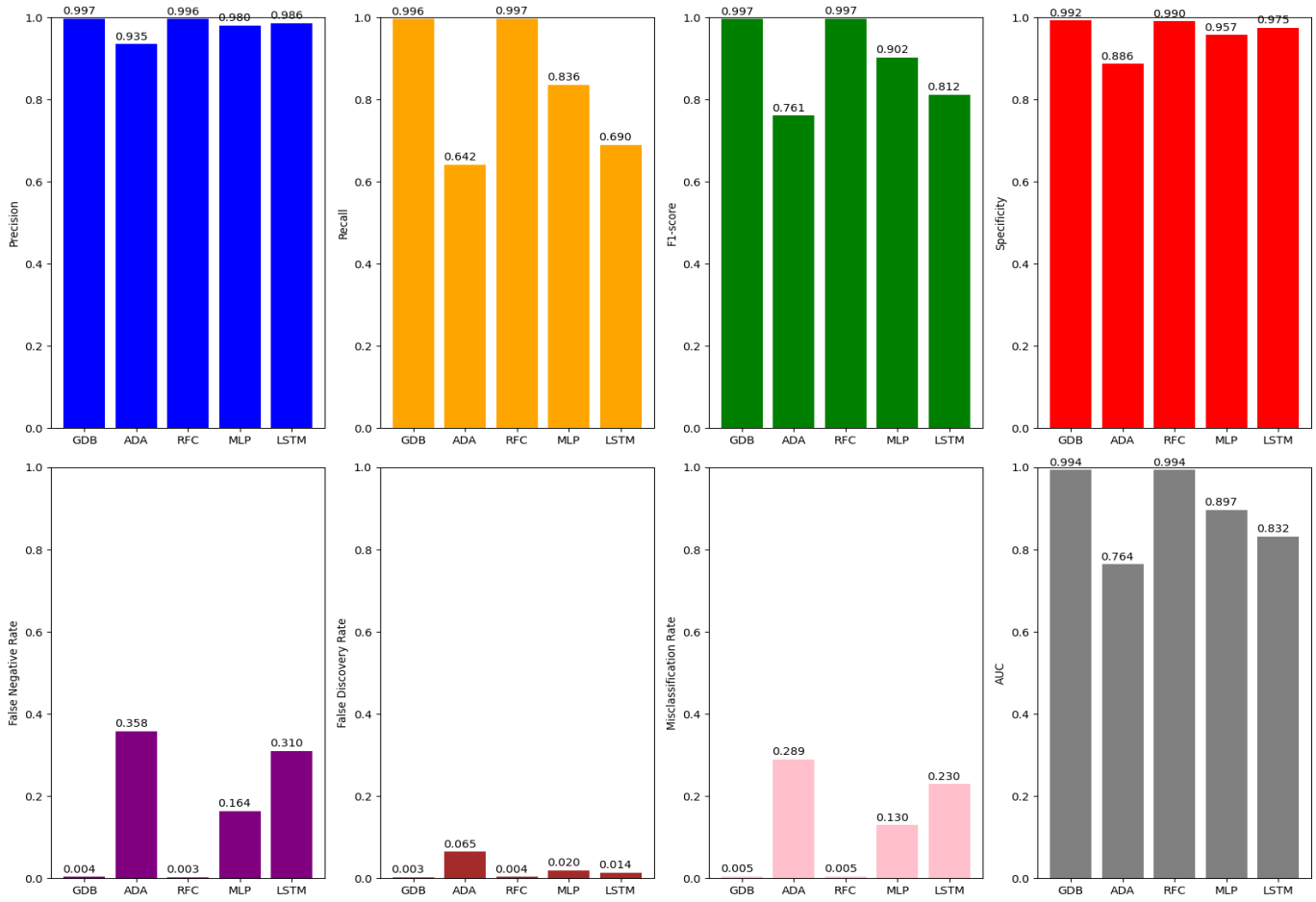
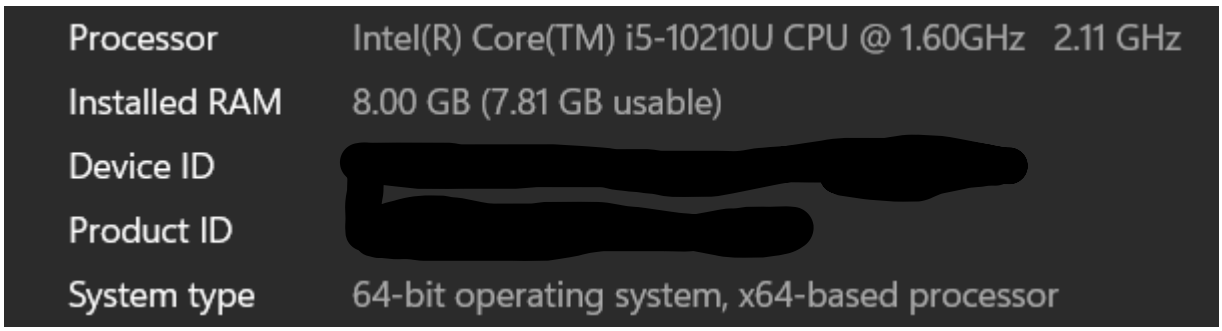


Fig 3.9 Evaluation Parameters Comparison

This analysis demonstrates that feature elimination effectively reduces computational costs without significantly compromising model performance. Consequently, the models become more efficient and scalable, which is particularly advantageous in real-time or resource-constrained environments.

All these results are evaluated on a system with following properties.



Chapter 4

Conclusions and Future Scope

This thesis aimed to develop and evaluate machine learning models for predicting liver disease, leveraging a dataset that includes a range of medical and biochemical features. By implementing various machine learning algorithms and optimizing them through feature elimination and data preprocessing techniques, this study sought to identify the most effective models for accurate and efficient liver disease prediction. The following sections provide a detailed summary of the dataset collection, preprocessing steps, model development, evaluation, and the impact of feature elimination on model performance and computational efficiency.

The initial phase of this thesis involved collecting relevant and high-quality datasets for liver disease prediction. A significant challenge in the dataset was the imbalance between the classes, with data instances for 'data with disease' significantly outnumbering those for 'data without disease'. To address this, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. SMOTE generated synthetic samples for the minority class, thereby balancing the dataset and enhancing the model's ability to predict liver disease accurately across all classes. This technique proved crucial in improving the model's robustness and predictive power.

The dataset was split into training and testing sets using a 60-40 ratio. This approach ensured that 60% of the data was used for training the models, while the remaining 40% was reserved for testing their performance. Splitting the data in this manner provided a substantial portion for model development and a significant portion for objective evaluation, ensuring the models' generalization capabilities were accurately assessed.

The core of this thesis involved developing and evaluating multiple machine learning models, categorized into ensemble methods, boosting methods, and neural networks. Each category was chosen for its unique strengths and capabilities in handling complex datasets.

Random Forest (RFC) achieved the highest test accuracy of 99.54%. As an ensemble method, it constructs a multitude of decision trees and merges their results to improve accuracy. By aggregating the predictions of

multiple trees, Random Forest demonstrated robustness and effectiveness in handling diverse data patterns, making it a highly reliable model for liver disease prediction.

Gradient Boosting (GB) and AdaBoost also showed strong performance. Gradient Boosting achieved a test accuracy of 99.52%, effectively capturing complex patterns through iterative error correction. AdaBoost, while achieving a lower test accuracy of 71.05%, highlighted the importance of boosting techniques in improving model accuracy by focusing on difficult-to-predict instances. However, AdaBoost's sensitivity to noisy data and outliers was noted as a limitation.

The Multi-Layer Perceptron (MLP) and Long Short-Term Memory (LSTM) networks demonstrated the ability to model complex relationships within the data. The MLP achieved a test accuracy of 87.00%, showing its capability to capture non-linear relationships, though requiring significant tuning of hyperparameters. The LSTM, typically used for sequence prediction, achieved a test accuracy of 77.00%, indicating that while it is powerful for sequential data, it was less suited for the non-sequential nature of this dataset.

Feature Elimination and Runtime Optimization To further enhance the models' efficiency, feature elimination was performed. By reducing the number of features from 10 to 5, significant reductions in runtime were observed without substantial compromise in performance metrics.

The runtime for GD Boost decreased by approximately 51%, from 6240.70 seconds to 3052.79 seconds. The performance metrics (Accuracy, Precision, Recall, F1-Score, AUC, False Negative Rate, and False Discovery Rate) remained relatively stable, with only a marginal decrease, demonstrating that feature elimination was an effective strategy for reducing computational cost while maintaining model performance.

The runtime for Random Forest decreased by approximately 47%, from 601.08 seconds to 316.43 seconds. Similar to GD Boost, the performance metrics remained stable, indicating that feature elimination did not significantly compromise the model's accuracy and robustness. This result underscores the advantage of feature elimination in making the models more efficient and scalable, particularly in real-time or resource-constrained environments.

In conclusion, this thesis demonstrated the effectiveness of machine learning models in predicting liver disease, with ensemble methods, particularly Random Forest, and boosting methods like Gradient Boosting, achieving the highest accuracy. The preprocessing steps, including normalization, label encoding, and SMOTE, were crucial in preparing the dataset for model training and ensuring robust performance.

The feature elimination process significantly reduced computational costs without substantially compromising model performance. This optimization is particularly beneficial for deploying models in real-time or resource-constrained environments, where efficiency is as critical as accuracy.

Overall, the results of this thesis highlight the potential of machine learning in improving diagnostic accuracy for liver disease, offering a reliable tool for early detection and timely medical intervention. Future work could explore further optimization techniques, the integration of additional data sources, and the application of these models to other medical conditions to enhance their generalizability and impact.

Future Directions

While this thesis has laid a strong foundation for using machine learning models in liver disease prediction, several areas warrant further exploration:

1. Hyperparameter Tuning: Although significant tuning was performed, exploring more advanced hyperparameter optimization techniques such as Bayesian optimization or evolutionary algorithms could further enhance model performance.

2. Feature Engineering: Investigating more sophisticated feature engineering techniques, including domain-specific transformations or the incorporation of additional features, could improve the models' predictive power.

3. Integration of Clinical Data: Combining biochemical data with clinical data, such as patient history and imaging results, could provide a more comprehensive dataset, potentially improving model accuracy and robustness.

4. Real-time Deployment: Developing a real-time prediction system integrated with electronic health records (EHR) systems could facilitate the practical application of these models in clinical settings, aiding healthcare professionals in making timely and accurate diagnoses.

5. Exploration of Other Algorithms: Exploring other advanced algorithms, such as deep learning models with more layers or hybrid models combining different techniques, could yield further improvements in prediction accuracy and efficiency.

6. Longitudinal Studies: Conducting longitudinal studies to assess the models' performance over time and their ability to adapt to new data could provide valuable insights into their long-term applicability and reliability.

By addressing these future directions, the work initiated in this thesis can be extended and refined, contributing to the ongoing advancement of machine learning applications in healthcare and ultimately improving patient outcomes.

References

[1] S. M. Ganie and P. K. Dutta Pramanik, "A comparative analysis of boosting algorithms for chronic liver disease prediction," *Healthcare Analytics*, vol. 5, 2024, doi: 10.1016/j.health.2024.100313.

[2] A. Q. Md, S. Kulkarni, C. J. Joshua, T. Vaichole, S. Mohan, and C. Iwendi, "Enhanced Preprocessing Approach Using Ensemble Machine Learning Algorithms for Detecting Liver Disease," *Biomedicines*, vol. 11, no. 2, 2023, doi: 10.3390/biomedicines11020581.

[3] R. Amin, R. Yasmin, S. Ruhi, M. H. Rahman, and M. S. Reza, "Prediction of chronic liver disease patients using integrated projection based statistical feature extraction with machine learning algorithms," *Inform Med Unlocked*, vol. 36, 2023, doi: 10.1016/j.imu.2022.101155.

[4] S. Dalal, E. M. Onyema, and A. Malik, "Hybrid XGBoost model with hyperparameter tuning for prediction of liver disease with better accuracy," *World J Gastroenterol*, vol. 28, no. 46, 2022, doi: 10.3748/wjg.v28.i46.6551.

[5] S. N. Pasha, D. Ramesh, S. Mohmmad, N. P., P. A. Kishan, and C. H. Sandeep, "Liver disease prediction using ML techniques," in *AIP Conference Proceedings*, 2022. doi: 10.1063/5.0081787.

[6] H. Luo *et al.*, "Factors influencing central lamina cribrosa depth: A multicenter study," *Invest Ophthalmol Vis Sci*, vol. 59, no. 6, 2018, doi: 10.1167/iovs.17-23456.

[7] A. Bayani *et al.*, "Identifying predictors of varices grading in patients with cirrhosis using ensemble learning," *Clin Chem Lab Med*, vol. 60, no. 12, 2022, doi: 10.1515/cclm-2022-0508.

[8] K. Thirunavukkarasu, A. S. Singh, M. Irfan, and A. Chowdhury, "Prediction of liver disease using classification Algorithms," in *2018 4th International Conference on Computing Communication and Automation, ICCCA 2018*, 2018. doi: 10.1109/CCAA.2018.8777655.

[9] L. Meng, W. Treem, G. A. Heap, and J. Chen, "A stacking ensemble machine learning model to predict alpha-1 antitrypsin deficiency-associated liver disease clinical outcomes based on UK Biobank data," *Sci Rep*, vol. 12, no. 1, 2022, doi: 10.1038/s41598-022-21389-9.

[10] V. Vats, L. Zhang, S. Chatterjee, S. Ahmed, E. Enziama, and K. Tepe, "A Comparative Analysis of Unsupervised Machine Techniques for Liver Disease Prediction," in *2018 IEEE International Symposium on Signal Processing and Information Technology, ISSPIT 2018*, 2018. doi: 10.1109/ISSPIT.2018.8642735.

[11] L. Alice Auxilia, “Accuracy Prediction Using Machine Learning Techniques for Indian Patient Liver Disease,” in *Proceedings of the 2nd International Conference on Trends in Electronics and Informatics, ICOEI 2018*, 2018. doi: 10.1109/ICOEI.2018.8553682.

[12] S. Kumar and S. Katyal, “Effective Analysis and Diagnosis of Liver Disorder by Data Mining,” in *Proceedings of the International Conference on Inventive Research in Computing Applications, ICIRCA 2018*, 2018. doi: 10.1109/ICIRCA.2018.8596817.

[13] P. Kumar and R. S. Thakur, “Diagnosis of Liver Disorder Using Fuzzy Adaptive and Neighbor Weighted K-NN Method for LFT Imbalanced Data,” in *6th IEEE International Conference on "Smart Structures and Systems", ICSSS 2019*, 2019. doi: 10.1109/ICSSS.2019.8882861.

[14] V. J. Gogi and M. M. Vijayalakshmi, “Prognosis of Liver Disease: Using Machine Learning Algorithms,” in *2018 International Conference on Recent Innovations in Electrical, Electronics and Communication Engineering, ICRIEECE 2018*, 2018. doi: 10.1109/ICRIEECE44171.2018.9008482.

[15] G. A. Shanbhag, K. A. Prabhu, N. V. S. Reddy, and B. A. Rao, “Prediction of Lung Cancer using Ensemble Classifiers,” in *Journal of Physics: Conference Series*, 2022. doi: 10.1088/1742-6596/2161/1/012007.

[16] E. N. Mutlu, A. Devim, A. A. Hameed, and A. Jamil, “Deep Learning for Liver Disease Prediction,” in *Communications in Computer and Information Science*, 2022. doi: 10.1007/978-3-031-04112-9_7.

[17] S. Afrin *et al.*, “Supervised machine learning based liver disease prediction approach with LASSO feature selection,” *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 6, 2021, doi: 10.11591/eei.v10i6.3242.

[18] M. Hassoon, M. S. Kouhi, M. Zomorodi-Moghadam, and M. Abdar, “Rule Optimization of Boosted C5.0 Classification Using Genetic Algorithm for Liver disease Prediction,” in *2017 International Conference on Computer and Applications, ICCA 2017*, 2017. doi: 10.1109/COMAPP.2017.8079783.

REFLECTION ON THE DESIGN PROCESS

The design process for this thesis began with clearly defining the objectives and outlining a structured work plan. The primary goal was to develop effective machine learning models for predicting liver disease using a comprehensive dataset. The process was divided into several phases: dataset collection, data preprocessing, model building and evaluation, and optimizing the models for performance and efficiency. This structured approach ensured that each phase was given adequate attention and resources, ultimately leading to a thorough and systematic investigation.

The design process for this thesis involved collecting high-quality datasets from reliable sources like the UCI Machine Learning Repository, emphasizing data integrity and completeness. Data preprocessing was intensive, including normalization, handling missing values, label encoding, and addressing class imbalance with SMOTE. These steps were crucial for preparing the data for effective model training, ensuring all features contributed equally to the learning process. Building and evaluating various machine learning models, including ensemble methods, boosting methods, and neural networks, revealed that Random Forest and Gradient Boosting were top performers with accuracy rates of 99.54% and 99.52%, respectively. Evaluating models using a multi-metric approach ensured a comprehensive assessment of their performance. Feature elimination significantly reduced runtime by about 50% for both Gradient Boosting and Random Forest without substantially compromising performance metrics, highlighting the importance of optimizing for both accuracy and computational efficiency in practical applications, especially in healthcare.

The design process for this thesis was marked by continuous learning and adaptation, where each phase provided valuable insights for iterative improvements. Challenges in data preprocessing underscored the need for rigorous data quality checks, which were diligently applied in the feature elimination phase. Iterative model evaluation refined selection criteria, enhancing the optimization of top-performing models. Collaboration and feedback from advisors and peers significantly enriched the research process, leading to refined methodologies and robust findings. The comprehensive journey, from dataset collection to model optimization, highlighted the effectiveness

of a structured and iterative approach. The successful identification of Random Forest and Gradient Boosting as the top models for liver disease prediction underscores the importance of thorough preparation, diverse methodologies, and flexibility in research design. These reflections emphasize the critical need for ongoing adaptation to achieve reliable outcomes in machine learning, particularly in complex fields like healthcare, and will inform future research endeavors, advancing machine learning applications in medical diagnosis.

APPENDICES

Appendix A: Datasets and Sources

- **Description of the datasets**

  This dataset comprises numerous features related to liver health and disease, providing a robust foundation for building predictive models. Key features include serum alkaline phosphatase (alkophos), albumin and globulin ratio (ag_ratio), total proteins (tp), serum glutamic-pyruvic transaminase (sgpt), and others, each playing a vital role in liver function assessment.

  To address the inherent class imbalance in the dataset—where data with liver disease (21,917 instances) significantly outnumbered data without liver disease (8,774 instances)—the Synthetic Minority Over-sampling Technique (SMOTE) was employed. SMOTE generated synthetic samples for the minority class, enhancing the dataset's balance and improving the model's ability to generalize across all classes.

  This comprehensive dataset, along with the rigorous preprocessing steps, laid a strong foundation for developing accurate and reliable machine learning models for liver disease prediction. The thorough preparation ensured that the models built were both robust and capable of providing meaningful insights into liver health.

- **Data preprocessing steps**

  Before applying any machine learning techniques, the dataset was meticulously verified for integrity and completeness. This verification process involved checking for missing values, ensuring consistent data types across all entries, and validating the data against known medical standards. Handling missing values was particularly critical, as incomplete data can skew model training and lead to inaccurate predictions. Normalization was applied to ensure that features with different scales, such as alkophos (typically around 150+) and ag_ratio (around 1 or less), contributed equally to the model training process.

  Furthermore, to address the inherent class imbalance in the dataset—where data with liver disease (21,917 instances) significantly outnumbered data without liver disease (8,774 instances)—the Synthetic Minority Over-sampling Technique (SMOTE) was employed. SMOTE generated synthetic samples for the minority class, enhancing the dataset's balance and improving the model's ability to generalize across all classes.

- **Link to datasets**

https://www.kaggle.com/datasets/abhi8923shriv/liver-disease-patient-dataset

SIMILARITY REPORT

Thesis_LiverDisease

ORIGINALITY REPORT

17%	10%	10%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Shahid Mohammad Ganie, Pijush Kanti Dutta Pramanik. "A comparative analysis of boosting algorithms for chronic liver disease prediction", Healthcare Analytics, 2024 Publication	2%
2	www.mdpi.com Internet Source	1%
3	Submitted to The University of Memphis Student Paper	1%
4	Shahid Mohammad Ganie, Pijush Kanti Dutta Pramanik, Zhongming Zhao. "Improved liver disease prediction from clinical data through an evaluation of ensemble learning approaches", BMC Medical Informatics and Decision Making, 2024 Publication	1%
5	Submitted to University of Surrey Student Paper	1%
6	fastercapital.com Internet Source	1%