Senior Design Project

(CSE4101)

Midterm Presentation

on

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

Supervisor

Prof.

DR. SAMRUDHI MOHDIWALE

Presented by - Group no. D1

HIMANSHU RANJAN (2041018156)

SUNEET KUMAR (2041011132)

HARSHIT MALVIA (2041013047)

RAHUL RAJ (2041018152)

Department of Computer Science & Information Technology

Faculty of Engineering & Technology(ITER),

Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar, Odisha

Presentation Outline

- **Introduction**
  - Overview
  - Problem(s) addressed
  - Motivations
- **Objectives**
- **Software, Hardware or Methods/Algorithms Specifications**
- **Benefits/Social Contribution**
- **Tentative roadmap**
- **Result and its analysis**
- **Conclusion**
- **References**

Introduction

Global burden of liver disease

1. High Mortality Rate: Liver disease contributes to two million deaths annually worldwide, necessitating predictive models to identify at-risk individuals for early intervention and prevention.

2. Diverse Causes: With causes ranging from viral hepatitis to alcohol and non-alcoholic fatty liver disease, predictive models help pinpoint specific risk factors, enabling tailored prevention strategies.

3. Emerging Trends: Addressing evolving trends like drug-induced liver injury requires predictive models to forecast risks accurately, aiding in medication management and prevention efforts.

.

Introduction

- **Overview**

1. **Enhancing Liver Disease Diagnosis with Machine Learning**:

- Leveraging Boosting Algorithms, Tree Classifications, LSTM networks, the study analyzes patient demographics and liver function metrics to classify individuals as diseased or healthy.

2. **Machine Learning's Impact on Healthcare Insights**:

- By evaluating accuracy, sensitivity, and specificity, the research uncovers vital clinical markers for liver disease diagnosis, promising improved care and reduced healthcare strain.

Introduction

- **Problem(s) addressed**
  - Global impact of liver diseases underscores the need for effective diagnostic approaches.
  - Early detection crucial for timely treatment initiation and improved patient outcomes.
  - Aim to leverage machine learning (ML) techniques for diagnosing liver diseases.
  - Utilize patient data including age, gender, and key liver function metrics for analysis.

Introduction

- **Motivations:**

1.Global Health Impact

- Liver diseases pose a significant health burden worldwide.
- Motivated by the pressing need to improve diagnosis and treatment outcomes for these conditions.

2.Early Intervention

- Early detection of liver diseases is critical for initiating timely interventions.
- Motivation stems from the potential to identify diseases in their early stages, when treatments are most effective.

3.Healthcare System Efficiency

- Optimizing diagnostic processes can lead to more efficient use of healthcare resources.
- Motivation lies in easing the burden on healthcare systems by streamlining diagnosis and treatment pathways.

Objective

- **Exploratory Data Analysis**

  Calculation of Mean, Median , Mode . Studying about deviation of the data and Efficient removal of null data present in the dataset.

- **To select the optimal features that increases the overall classification result**

  Use of Filter and Wrapper methods to access feature relevance.

  Use of RFE to eliminate unnecessary features.

- **To explore the deep learning models**

  Explore boosting algorithms, Random Forest Tree algorithms, and complex models including MLP and LSTM to efficiently capture sequential pattern for feature classification.

Software, Hardware or Methods/Algorithms Specifications

- **1.Statistics and Visualization**
- - Statistics: Employ metrics such as accuracy, precision, recall, F1-score, ROC curves, AUC, and confusion matrices to comprehensively evaluate model performance.
- - Visualization: Utilize libraries like Matplotlib, Seaborn, and Plotly to create visually informative representations of evaluation metrics, aiding in the clear interpretation of model performance.

- **2. Feature Selection Technique**
- - Filter Method: Independently assess feature relevance using predefined criteria such as correlation or statistical tests.
- - Wrapper Method: Iterate through feature subsets using model performance as a guide to identify the optimal subset.
- - Rfe – Recursive Feature Elimination

- **3. Deep Learning Methods**
- - Long Short-Term Memory (LSTM): Capture long-term dependencies in sequential patient data, enabling the model to learn complex temporal patterns relevant to liver disease diagnosis.
- - MLP: Multilayer Perceptron, a neural network model for complex pattern recognition and prediction tasks.

Dataset Description

1. **TB (Total Bilirubin):** Measures all bilirubin in the blood; high levels can indicate liver or bile duct issues.

- 2. **DB (Direct Bilirubin):** Measures processed bilirubin; high levels may suggest liver dysfunction or bile duct obstruction.

3. **Alkphos (Alkaline Phosphatase):** An enzyme indicating liver, bile duct, or bone disorders when elevated.

4. **Sgpt/ALT (Alanine Aminotransferase):** An enzyme; high levels often indicate liver damage.

5. **Sgot/AST (Aspartate Aminotransferase):** An enzyme; elevated levels suggest liver damage or muscle injury.

6. **TP (Total Protein):** Total protein in the blood; low levels can indicate liver or kidney disease, or malnutrition.

7**. ALB (Albumin):** Main protein from the liver; low levels can indicate liver or kidney disease, or malnutrition.

8. **A/G Ratio (Albumin/Globulin Ratio):** Ratio of albumin to globulin; abnormal ratios can indicate liver or kidney disease, or immune disorders.

These tests are used to assess liver health, detect liver or bile duct disorders, evaluate protein levels, and monitor for conditions like liver damage, bile duct obstruction, and malnutrition

FLOWCHART

Data Collection :

Data Count: 30,691 rows

Features : 10

Data Pre-processing :

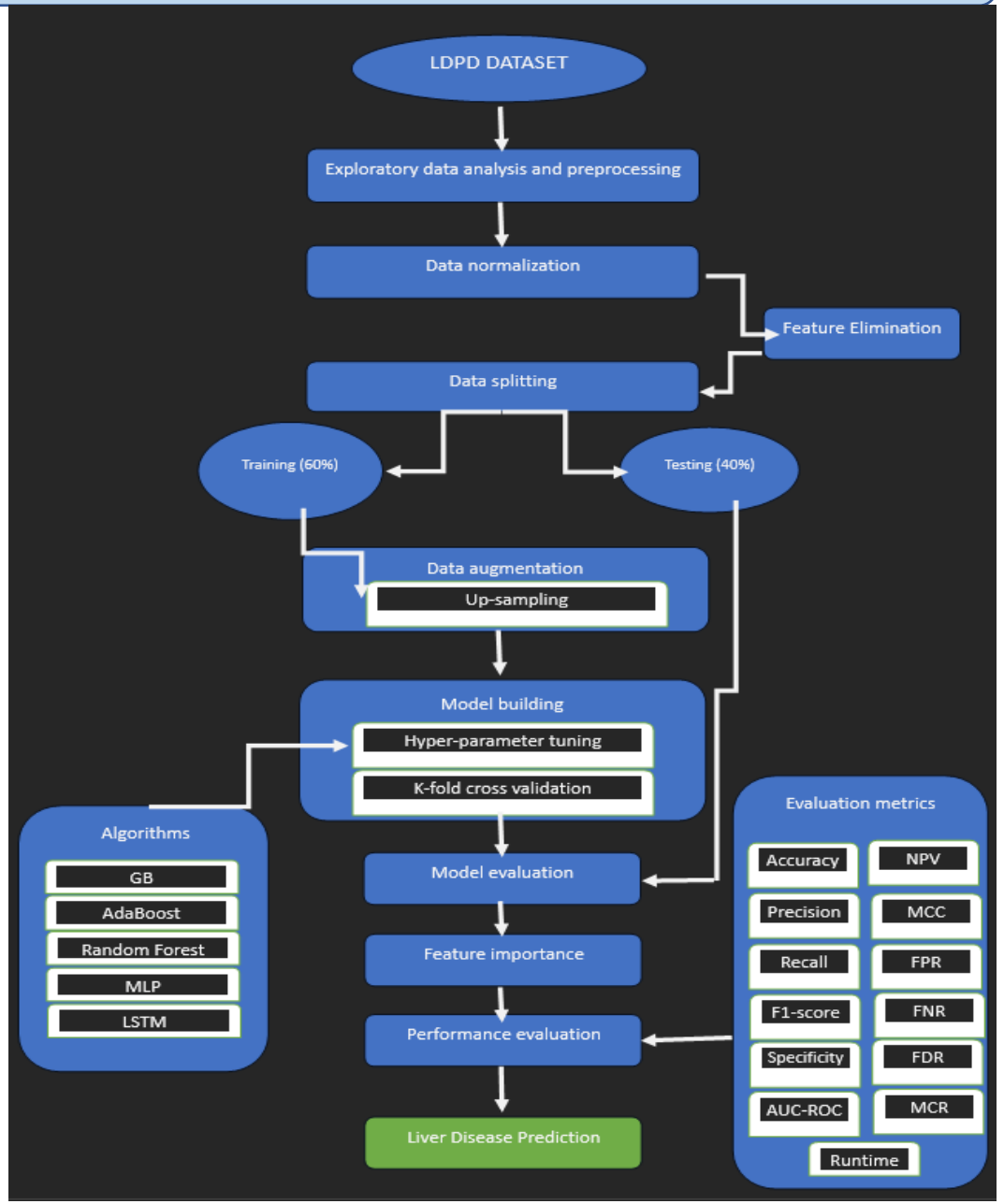
encompasses cleaning, transforming, and organizing raw data to enhance quality and usability. Steps include handling missing values, outlier detection, and feature engineering, crucial for effective analysis and modeling.

Data Normalization (used min-max scalar) :

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.

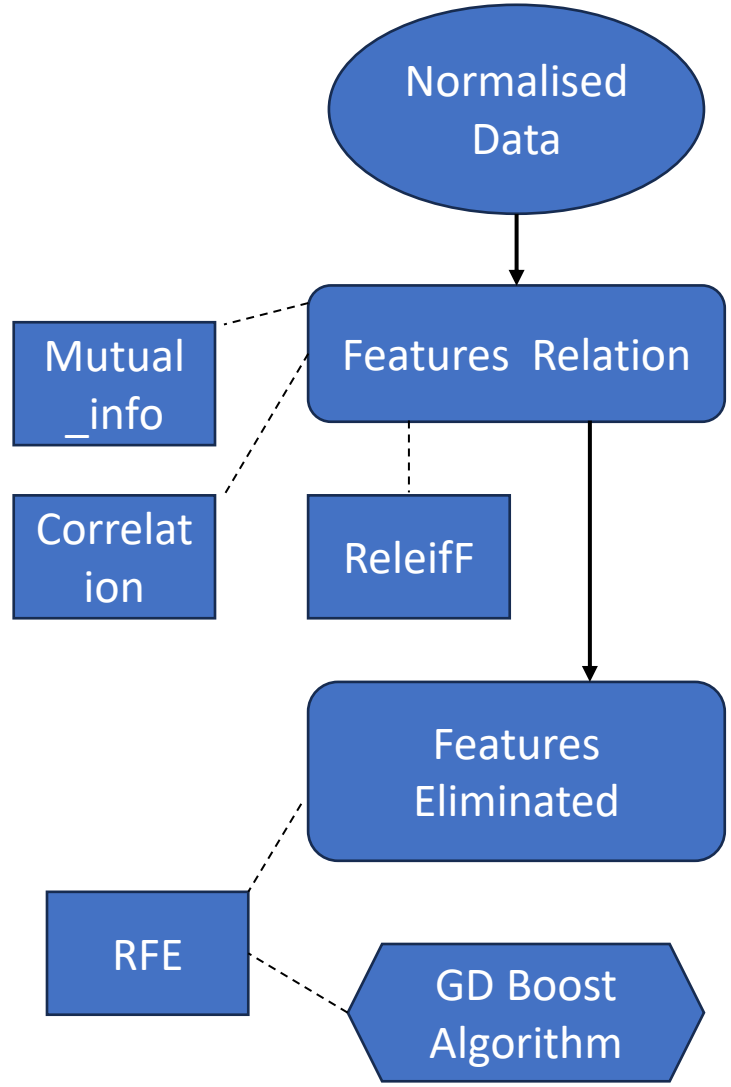


FLOWCHART (cont)

Feature Classification and Elimination -

Used different Classification Techniques to thoroughly verify the best features and finally Eliminated unwanted features using RFE.

- - mutual_info_classif: A feature selection method measuring the dependency between features and target variable in classification tasks, based on mutual information theory, aiding in identifying relevant features.
- - corr: Short for correlation, a statistical measure quantifying the strength and direction of the linear relationship between two variables, often used in feature selection to identify highly correlated features.
- - ReliefF: An algorithm for feature selection, focusing on finding features that are most discriminative for classification tasks by iteratively updating feature weights based on nearest neighbors.
- - RFE (Recursive Feature Elimination): A feature selection technique that recursively removes features from a model, ranking them by their impact on model performance, until the desired number of features is reached, aiding in identifying the most informative features.



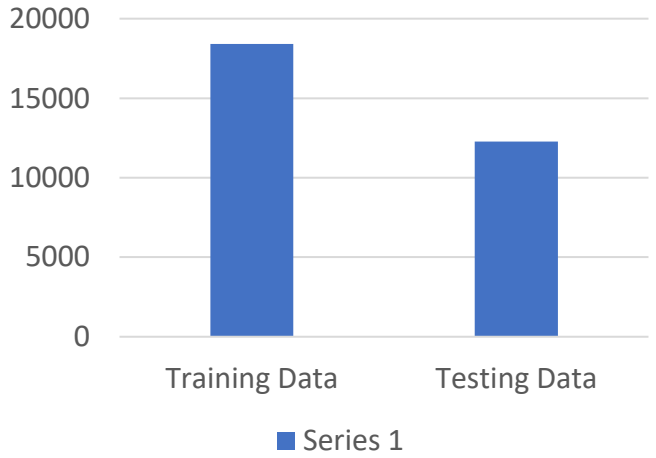
FLOWCHART (cont)

Data Splitting:

Data splitting involves partitioning a dataset into separate subsets, typically training and testing sets, to evaluate model performance. Ratio used is 60-40.

Training data - 18,414 rows

Testing data - 12,277 rows

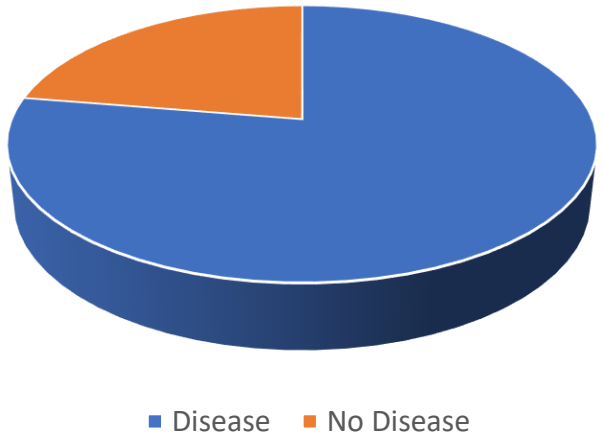


Upsampling using SMOTE:

Data with Disease: 21917

Data without Disease: 8774

SMOTE (Synthetic Minority Over-sampling Technique) is a method for addressing class imbalance by generating synthetic samples of the minority class, enhancing training data and improving model performance.



Model Building

Hyper-parameter Tuning : Hyper-parameter tuning is the process of optimizing the parameters of a machine learning model to improve its performance. Techniques such as grid search or random search are used to systematically explore the hyper-parameter space and find the best combination.

K-fold Cross-validation : K-fold cross-validation is a technique used to assess the performance of a machine learning model. The dataset is divided into k subsets, and the model is trained and evaluated k times, with each subset used as the validation data once, providing robust performance estimation.

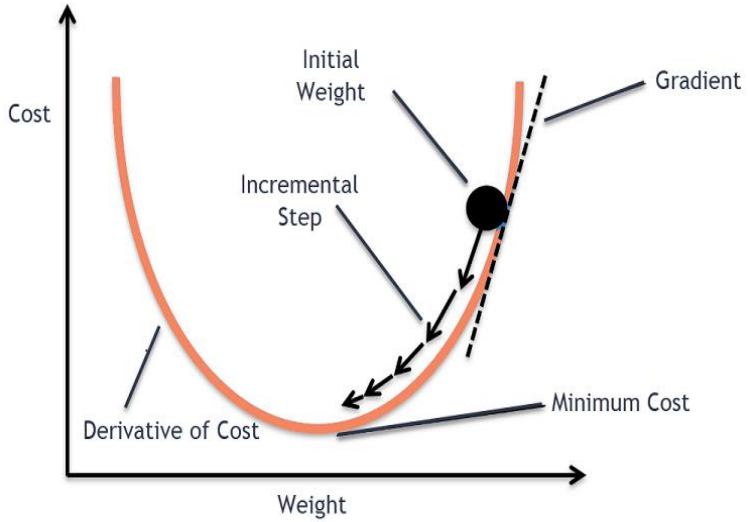
Algorithms:

➢ **Gradient Boosting (GB) Classifier:**

Working -

Gradient Boosting is an ensemble method that combines weak learners sequentially, correcting errors iteratively.

It optimizes by fitting new models to the residuals of previous ones, iteratively minimizing the loss function until convergence.



Benefits:

1. High predictive accuracy: GB combines multiple weak learners to create a strong predictive model, often achieving high accuracy.

2. Handles complex relationships: GB can capture complex relationships between features and target variables, making it suitable for predicting liver diseases.

3. Robust to overfitting: GB uses boosting, which sequentially improves model performance by focusing on hard-to-predict instances, reducing overfitting.

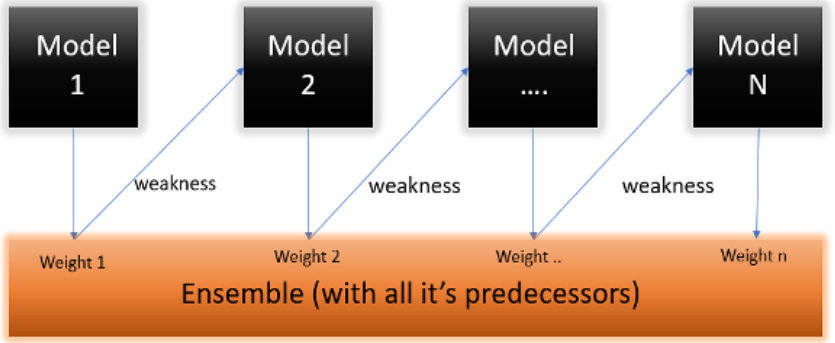
Algorithms:

➢AdaBoost Classifier:

Working:

AdaBoost sequentially trains weak classifiers, emphasizing misclassified instances in each iteration.

It assigns higher weights to misclassified instances, improving model focus on difficult examples and combining weak learners' predictions for a strong classifier.



Benefits:

1. Simple and easy to implement: AdaBoost is straightforward to implement and often requires minimal parameter tuning.

2. Adaptable to various classifiers: AdaBoost can be used with different base classifiers, allowing flexibility in model selection.

3. Focuses on hard examples: AdaBoost assigns higher weights to misclassified instances, enabling the model to focus on difficult-to-predict cases.

Algorithms:

➢Random Forest Classifier:

Working:

RFC utilizes an ensemble of decision trees, each trained on a random subset of features and instances.

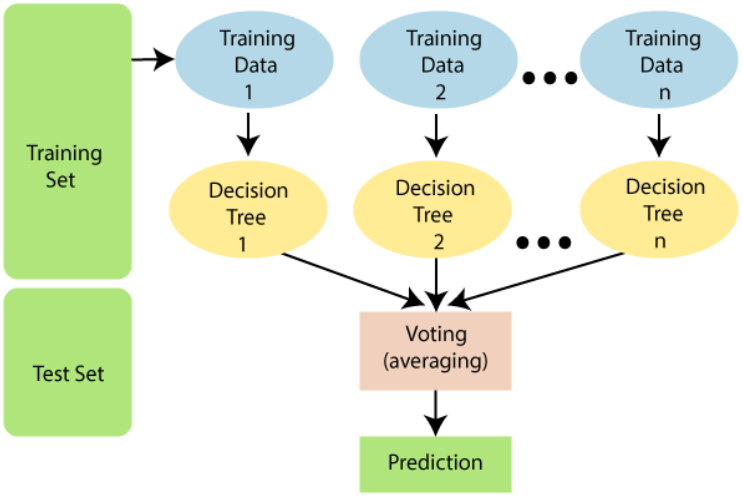
It aggregates predictions from individual trees to make final classifications, resulting in robust performance and reduced risk of overfitting.

Benefits:

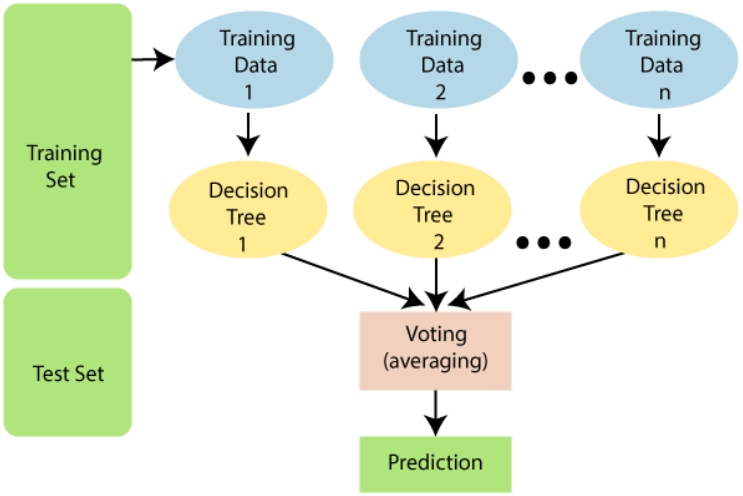
1. High predictive accuracy: Random Forest combines multiple decision trees to create an ensemble model with high predictive performance.

2. Robust to overfitting: Random Forest mitigates overfitting by averaging predictions from multiple trees and bootstrapping training data.

3. Handles high-dimensional data: Random Forest can handle datasets with a large number of features without feature selection.



Algorithms:



➢Multilayer Perceptron (MLP):

Working:

MLP is a type of artificial neural network with multiple layers, including input, hidden, and output layers.

It learns by iteratively adjusting weights through backpropagation of errors, aiming to minimize the difference between predicted and actual outputs.

Benefits:

1. Ability to capture complex patterns: MLPs can learn intricate patterns and relationships in data, making them suitable for complex prediction tasks like liver disease prediction.

2. Adaptability to various data types: MLPs can handle different types of data, including numerical, categorical, and textual data.

3. Flexibility in architecture: MLPs allow customization of network architecture, including the number of layers, neurons, and activation functions.

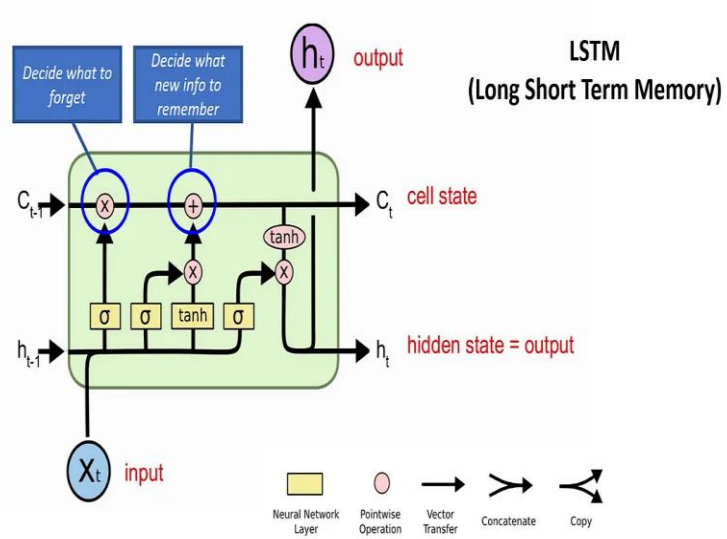
Algorithms:

➢Long Short-Term Memory (LSTM):

Working:

LSTM (Long Short-Term Memory) networks have specialized memory cells that can retain information over time, enabling them to learn long-term dependencies in sequential data.

Through gates such as input, forget, and output gates, LSTM cells regulate the flow of information, allowing them to selectively remember or forget past information and control the information flow to the next time step, enhancing their ability to capture complex temporal patterns.



Benefits:

1. Sequential data processing: LSTM networks are well-suited for sequential data processing, making them ideal for time series data like liver disease progression.

2. Memory retention: LSTMs can retain information over long sequences, capturing long-term dependencies in data.

3. Handle variable-length sequences: LSTMs can process sequences of varying lengths, enabling them to handle diverse temporal data.

Result and its analysis

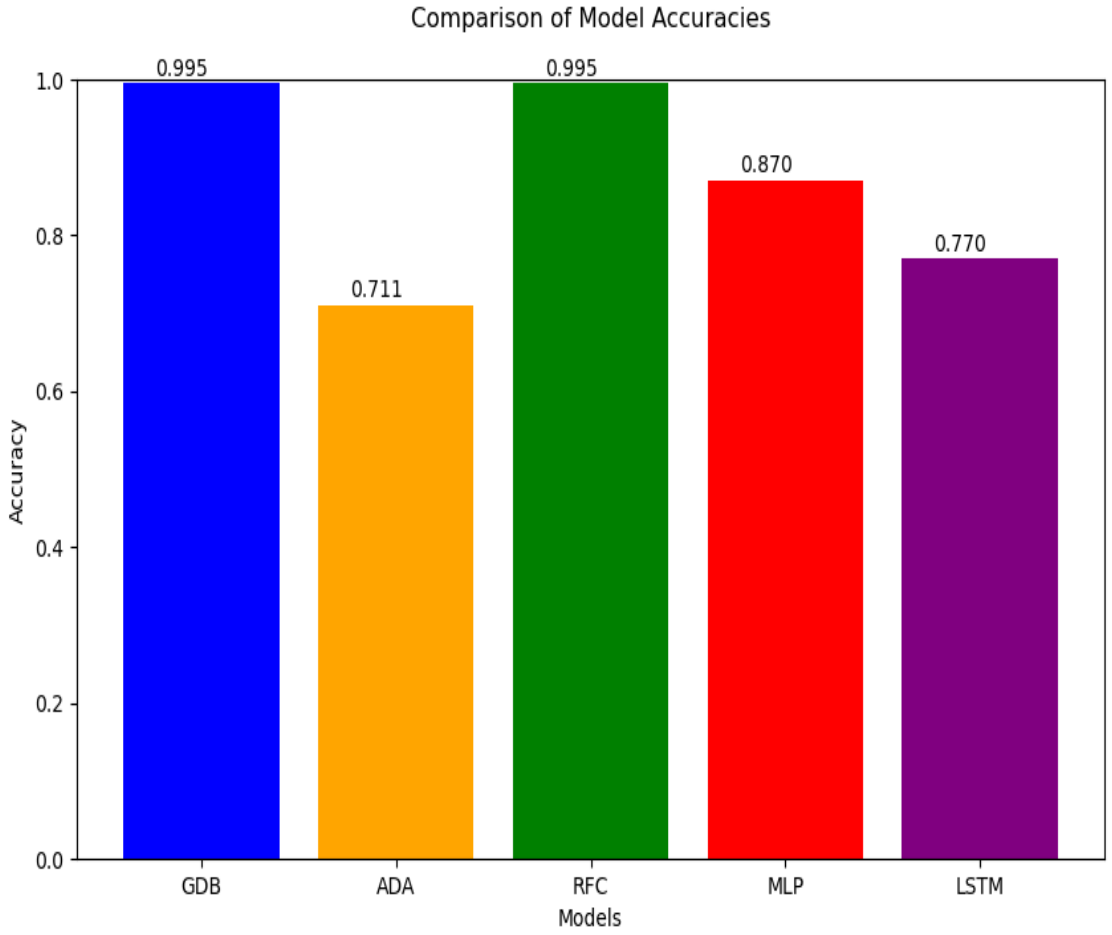
Based on all the models accuracy –

1. Gradient Boosting (gdb) and Random Forest Classifier (rfc) achieve the highest accuracy at 99.5%, indicating their effectiveness in capturing complex patterns.

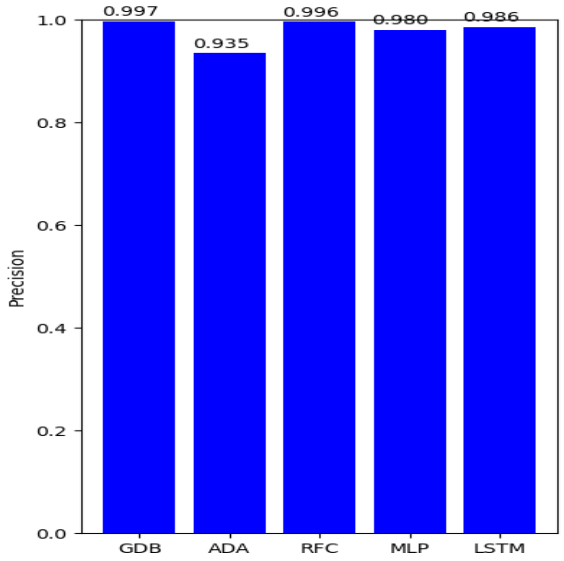
2. Multilayer Perceptron (mlp) follows with an accuracy of 87%, showcasing its proficiency in learning intricate data relationships.

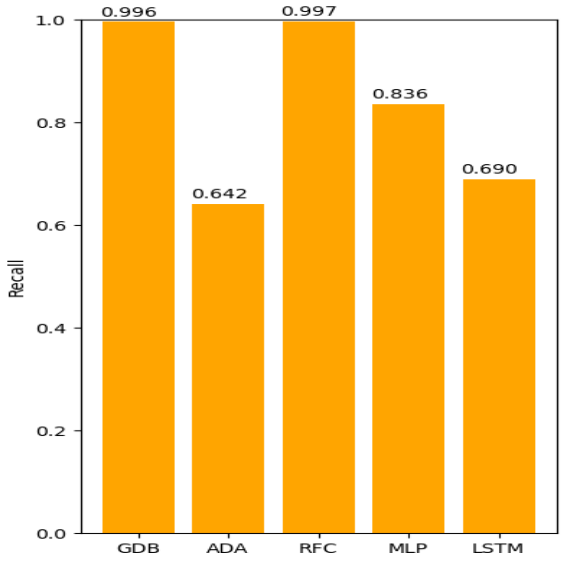
3. Adaptive Boosting (ada) and Long Short-Term Memory (LSTM) models exhibit comparatively lower accuracies at 71.1% and 77% respectively, suggesting potential challenges in capturing underlying patterns.

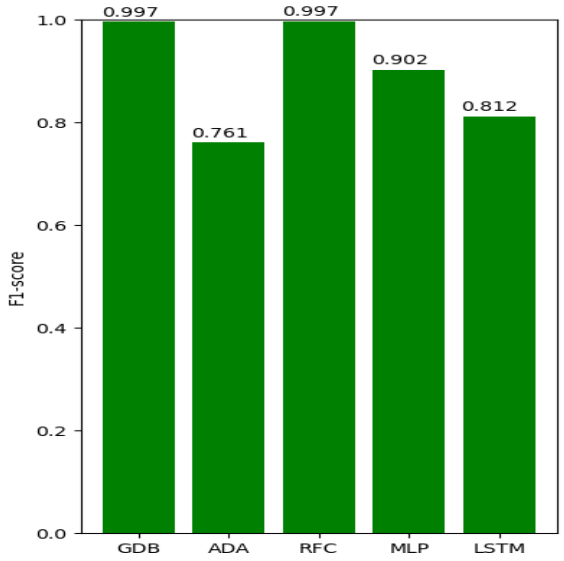
4. Despite varying accuracies, each model contributes uniquely, and considerations like computational cost and interpretability should guide model selection.

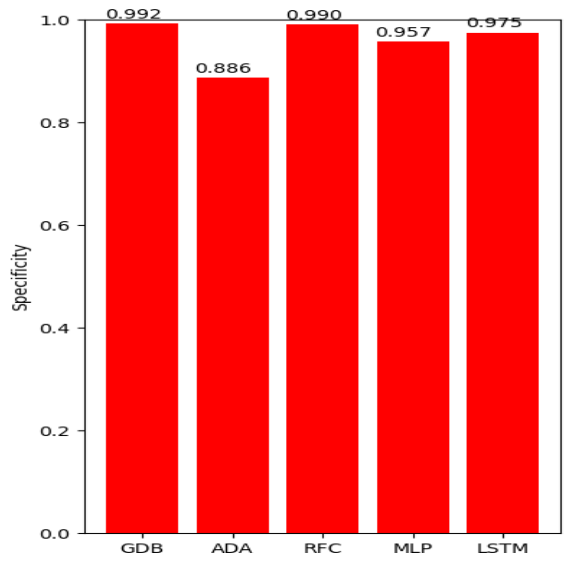


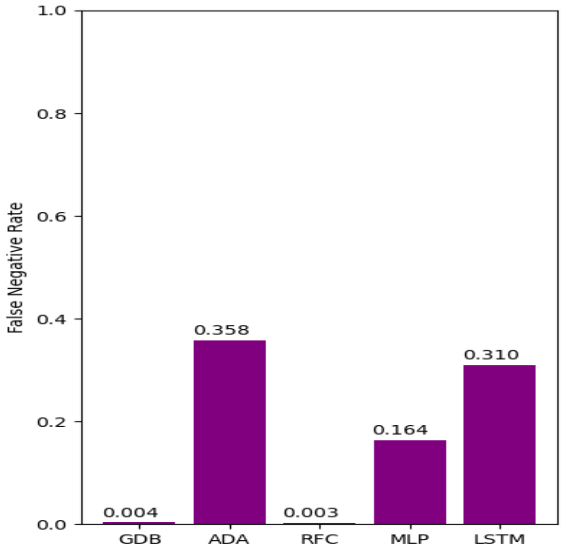
Result and its analysis

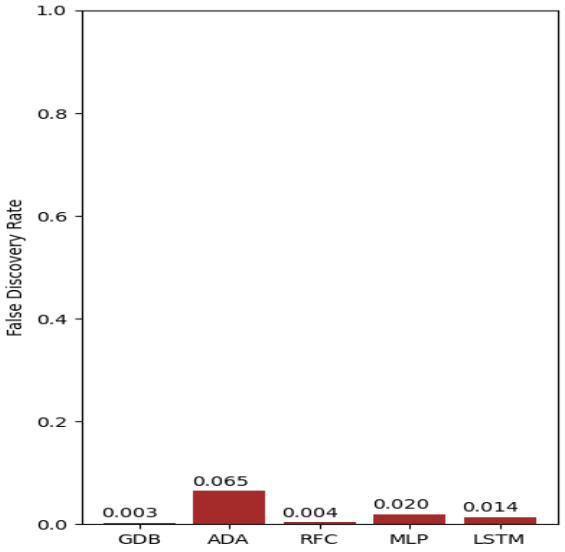


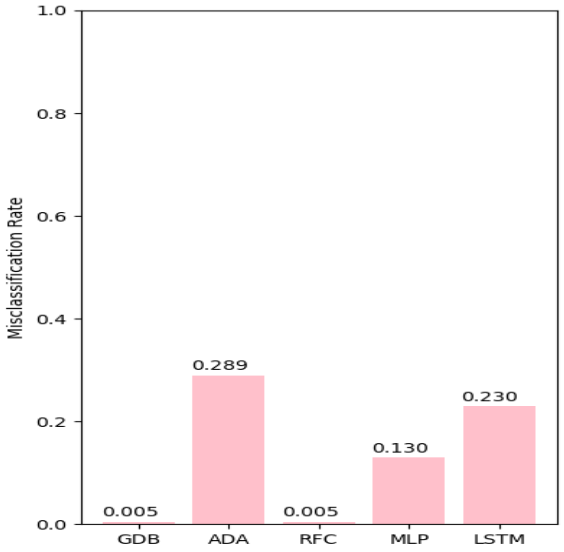


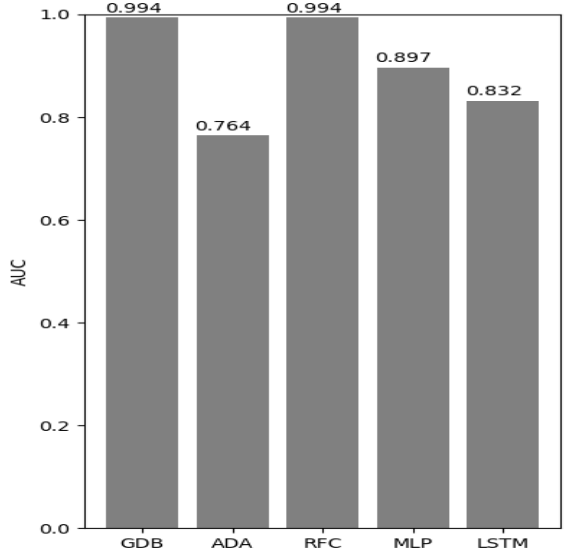












Result and its analysis (cont)

Based on F1-Score:

F1-score is the harmonic mean of precision and recall. It provides a balance between precision and recall.

1. Gradient Boosting (gdb) and Random Forest Classifier (rfc) boast high F1 scores of 0.997, indicating their robust performance in achieving a balance between precision and recall, essential for classification tasks.

2. Multilayer Perceptron (mlp) follows closely with an F1 score of 0.902, showcasing its effectiveness in capturing both precision and recall, which is crucial for tasks with imbalanced classes or complex decision boundaries.

These scores highlight the strengths of each model in accurately classifying instances while considering both false positives and false negatives.

Benefits/Social Contribution

1. Improved Healthcare Outcomes

- By enhancing the accuracy and efficiency of liver disease diagnosis, this research can lead to earlier detection and treatment initiation, ultimately improving patient outcomes and quality of life.

2. Reduced Healthcare Costs

- Timely diagnosis facilitated by advanced machine learning techniques can potentially reduce the economic burden associated with liver diseases. Early interventions may mitigate the need for expensive treatments or hospitalizations, leading to cost savings for healthcare systems and patients alike.

3. Accessible Healthcare Solutions

- By leveraging machine learning algorithms, particularly those capable of processing large datasets and complex patterns, this research can contribute to the development of accessible and scalable healthcare solutions. These advancements have the potential to benefit communities worldwide, especially in regions with limited access to specialized medical expertise.

References

- Amin, R., Yasmin, R., Ruhi, S., Rahman, M. H., & Reza, M. S. (2023). Prediction of chronic liver disease patients using integrated projection based statistical feature extraction with machine learning algorithms. *Informatics in Medicine Unlocked*, *36*, 101155.
- Kuzhippallil, M. A., Joseph, C., & Kannan, A. (2020, March). Comparative analysis of machine learning techniques for indian liver disease patients. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)* (pp. 778-782). IEEE.
- Tian, Y., Liu, M., Sun, Y., & Fu, S. (2023). When liver disease diagnosis encounters deep learning: Analysis, challenges, and prospects. *iLIVER*, *2*(1), 73-87.
- D'Amico, G., Colli, A., Malizia, G., & Casazza, G. (2023). The potential role of machine learning in modelling advanced chronic liver disease. *Digestive and Liver Disease*, *55*(6), 704-713.
- Ganie, S. M., & Pramanik, P. K. D. (2024). A comparative analysis of boosting algorithms for chronic liver disease prediction. *Healthcare Analytics*, 100313.
- Singh, J., Bagga, S., & Kaur, R. (2020). Software-based prediction of liver disease with feature selection and classification techniques. *Procedia Computer Science*, *167*, 1970-1980.

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%

THANK YOU