TOWARDS EQUITABLE AND TRANSPARENT DATA-DRIVEN MODELS IN HEALTHCARE BY INVESTIGATING FAIRNESS AND EXPLAINABILITY

A Project Report Submitted

in Partial Fulfilment of the Requirements

for the Degree of

Bachelors of Technology(Hon)

in

Computer Science and Engineering

by

Konyala Anish Reddy

(Roll No. 2021BCS0099)

to

DEPARTEMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY KOTTAYAM-686635, INDIA

April 2024

DECLARATION

I, **Konyala Anish Reddy** (**Roll No: 2021BCS0099**), hereby declare that, this report entitled **"Towards Equitable and Transparent Data-Driven Models in Healthcare by investigating Fairness and Explainability"** submitted to Indian Institute of Information Technology Kottayam towards partial requirement of **Bachelor of Technology(Hon)** in **Computer Science and Engineering** is an original work carried out by me under the supervision of **Dr. Ebin Deni Raj** and has not formed the basis for the award of any degree or diploma, in this or any other institution or university. I have sincerely tried to uphold the academic ethics and honesty. Whenever an external information or statement or result is used then, that have been duly acknowledged and cited.

Kottayam-686635 **Konyala Anish Reddy**

April 2024

CERTIFICATE

This is to certify that the work contained in this project report entitled **"Towards Equitable and Transparent Data-Driven Models in Healthcare by investigating Fairness and Explainability"** submitted by **Konyala Anish Reddy** (**Roll No: 2021BCS0099**) to Indian Institute of Information Technology Kottayam towards partial requirement of **Bachelor of Technology(Hon)** in **Computer Science and Engineering** has been carried out by him under my supervision and that it has not been submitted elsewhere for the award of any degree.

Kottayam-686635
April 2024

Dr. Ebin Deni Raj
Project Supervisor

ABSTRACT

In this project we identify the causal relationships within data, understanding the impact of one variable on another. The project uses a combination of statistical techniques and machine learning algorithms to identify and interpret causal connections within tabular data. It focuses on uncovering hidden cause-and-effect relationships in datasets like observational healthcare records paticularly in liver dataets. In this project, we employ advanced Explainable Artificial Intelligence (XAI) techniques to infer and explain causal relationships within tabular data. Using advanced Explainable AI (XAI) techniques enhances our ability to interpret and understand the inferred causal relationships in the data. We use SHAP and LIME method to find the relationships between the data. The aim is to determine how changes in input variables (features) would impact the outcome, providing a better understanding of causal dependencies.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

As AI systems make decisions that impact individuals and societies, there's a growing demand for understanding how these decisions are made for accountability and ensuring ethical AI deployment. The most successful methods are so complex that it is difficult for a human to re-trace, to understand, and to interpret how a certain result was achieved. Explainability, interpretability, and understandability are therefore driven by the opaque nature of these "black-box" methodologies in machine learning and artificial intelligence. The inadequate explainability of machine learning limits the widespread implementation of AI in healthcare applications where decision-makers need an explanation of the reasoning behind the decisions. If AI cannot explain itself in the domain of healthcare, then its risk of making a wrong decision may override its advantages of accuracy, speed and decision-making efficiency.

Chapter 2

Literature Review

The literature review explores Explainable Artificial Intelligence (XAI) in healthcare. XAI helps us understand how AI makes decisions by providing clear explanations. For instance, in medicine, doctors need more than just a yes or no answer from AI—they need to know why AI makes certain predictions to support their diagnosis.

Various methods for comprehending AI's role in healthcare are included in the review. These include machine learning algorithms, statistical techniques, and specially created, easily understood models. While each method has advantages and disadvantages, when combined, they aid medical practitioners in making more informed decisions based on AI predictions.

One exciting area discussed is Counterfactual Explanations, where we simulate different scenarios to see how changes might affect patient outcomes. This can result in improved therapies and care plans by assisting medical professionals in understanding why AI makes certain predictions. In summary, the review demonstrates how XAI is revolutionizing healthcare by improving the understandability and trustworthiness of AI for both physicians and patients.

Literature Review is divided into three parts:

1. EXplainable AI in Healthcare
2. Unmasking Dependencies Using EXplainable AI
3. Counterfactual Explanations in Healthcare

2.1 Explainable AI in Healthcare

Explainable AI is a branch of AI which deals with the interpretability and transparency of AI models and their decisions. It promotes a set of tools, techniques, and algorithms that can generate high-quality interpretable, intuitive, human-understandable explanations of AI decisions. Explanations supporting the output of a model are crucial, e.g., in precision medicine, where experts require far more information from the model than a simple binary prediction for supporting their diagnosis.

The papers presented cover various aspects of Explainable Artificial Intelligence (XAI) in healthcare. Dave et al.[1] concentrate on the analysis of cardiovascular health, employing Bayesian, XGBoost, and SVM techniques to create predictive models for illness detection and treatment. In order to develop precision medicine techniques, Kim et al.[2] investigate breast cancer detection and surgical recommendations using datasets like the Wisconsin (Diagnostic) dataset and Tomography image dataset (LUNA16) in conjunction with algorithms like SVM and XGBoost. Using techniques like SHAP and LIME, Smith et al. [3] address COVID-19 prognostic forecasts with the goal of facilitating prompt interventions and resource allocation. The strategies Johnson et al.[4] probably investigate could help improve model performance and interpretability by resolving class imbalance in healthcare datasets. When taken as a whole, these studies enhance patient outcomes, guide clinician judgment, and progress healthcare management strategies through the application of XAI techniques.

2.2 Unmasking Dependencies Using Explainable AI

Here are the different XAI techniques for unmasking dependencies:

2.2.1 Statistical Techniques

The use of numerous statistical techniques, including regression analysis, correlation analysis, and causal inference algorithms, is a part of statistical techniques in the healthcare industry. In order to examine medical records and find correlations as well as possible causative relationships between variables, these approaches are used in conjunction with approaches such as Propensity Score Matching and Instrumental Variables.

The capacity of statistical techniques to yield quantitative estimates of causality and association is one of its main benefits. Researchers can quantify the correlations between various aspects in healthcare data by using rigorous statistical approaches. This allows researchers to acquire insights into the underlying mechanisms and contributing factors that affect patient outcomes.

It's crucial to remember that statistical techniques could have some drawbacks. For example, they might not fully capture intricate interactions between variables and frequently rely on strong assumptions about data distribution. Additionally, potential confounding variables and biases in observational data must be carefully taken into account when interpreting the results of statistical studies. Notwithstanding these drawbacks, statistical techniques are essential for evaluating medical data and providing guidance for evidence-based choices in clinical practice and public policy.

2.2.2 Machine Learning Algorithms

Machine learning algorithms, such as Random Forests, Decision Trees, and Bayesian Networks, are essential for identifying complex causal linkages and dependencies in tab-

ular healthcare data. Large dataset analysis is an area in which these algorithms shine, and they are skilled at identifying intricate relationships between variables. These models can uncover important contributing elements through feature importance analysis, which helps to clarify the main drivers behind different healthcare outcomes.

The capacity of machine learning algorithms to identify nonlinear relationships—which conventional statistical methods could miss—is one of its main benefits in the healthcare industry. They also show efficiency in managing large volumes of data, which makes them very useful in the healthcare industry where datasets can be large and complex.

But problems like interpretability problems and overfitting—particularly with high-dimensional data—are significant concerns. As models become more intricate, there is a chance that they could overfit to data noise, which could undermine the findings' generalizability. Moreover, black-box models—like deep neural networks—may be difficult to read, which makes it difficult to comprehend the logic underlying the model's predictions—a critical skill for earning respect and acceptance in therapeutic settings.

For machine learning algorithms to be applied in healthcare effectively, interpretability and model complexity must be balanced. These difficulties can be lessened by employing strategies like model regularization and using interpretable models in addition to more complicated ones, guaranteeing that the conclusions drawn from machine learning are precise and useful in real-world healthcare scenarios.

2.2.3 Interpretable Machine Learning Models

Essential elements of Explainable Artificial Intelligence (XAI) are Interpretable Machine Learning Models, which provide visible insights into how input features affect outcomes in healthcare datasets. For this, models like Decision Trees, Logistic Regression, and Linear Regression are frequently used. These models make decision-making processes easier to grasp by giving clear explanations of the correlations between input factors

and predictions. This makes them especially helpful for stakeholders who might not be machine learning experts.

Interpretable Machine Learning Models are important in healthcare applications because of their simplicity and ease of comprehension. Collaboration between data scientists and domain specialists is facilitated by the ease with which policymakers and healthcare professionals can understand the elements driving predictions or judgments. They also build trust.

Nevertheless, despite their benefits, Interpretable Machine Learning Models might not be able to fully capture the intricate correlations found in healthcare data. Their tree-based or linear structures could find it difficult to represent complex relationships or non-linear connections, which could lead to the omission of minute but important patterns. Because of this, even while these models offer insightful information, it's possible that they don't completely utilize the wealth of data found in intricate healthcare databases.

All things considered, Interpretable Machine Learning Models are an essential component of XAI in healthcare, providing clarity and openness in the decision-making process. Even though they might not be able to fully capture all the subtleties in the data, their ability to help healthcare stakeholders and data scientists communicate and understand one another is crucial for producing insightful discoveries and better patient care.

2.2.4 Counterfactual Explanations

Explainable Artificial Intelligence (XAI) explanations provide a persuasive way to comprehend causal relationships in healthcare datasets. Through the simulation of modifications to input variables and the observation of their effects on results, this method offers important insights into the ways in which changes to certain components may affect

patient health. In the healthcare industry, where causality is frequently intricate and multidimensional, Counterfactual Explanations provide a sophisticated comprehension of the connections between diverse circumstances and patient outcomes.

The capacity to provide practical insights for healthcare decision-making is the strength of Counterfactual Explanations. Healthcare professionals can investigate potential interventions or modifications to treatment procedures through simulated situations, enabling better informed and customized care plans. This strategy has the potential to improve patient outcomes by promoting evidence-based decision-making and allowing for the customization of interventions for each patient needs.

However, the effective utilization of Counterfactual Explanations requires careful consideration of various factors. From selecting appropriate variables to perturb to evaluating the ethical implications of simulated changes, healthcare professionals must navigate complexities to ensure responsible and effective use of this technique. While Counterfactual Explanations offer significant potential for enhancing decision-making in healthcare, their application necessitates thoughtful consideration of clinical context, ethical considerations, and analytical techniques to maximize their benefits.

2.2.5 Model-Agnostic Techniques

A key component of XAI is Model-Agnostic Techniques, which use approaches such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations). Regardless of the complexity of the machine learning model, these methods are intended to explain its predictions. Understanding the inner workings of machine learning models applied to healthcare data is made possible by Model-Agnostic Techniques, which reveal causal relationships and offer insights into the behavior of complicated models.

The capacity to provide causal link insights regardless of the underlying machine

learning model utilized in healthcare data analysis is a key benefit of model-agnostic techniques. Their versatility and applicability to different machine learning models stem from their lack of reliance on particular methods. Furthermore, these methods offer both regional and worldwide explanations, enabling users to understand model predictions at both individual instance and overall dataset levels.

However, Model-Agnostic Techniques may have limitations. They may not capture interactions between features accurately, potentially overlooking important nuances in the data. Moreover, there might be a computational overhead associated with these techniques, especially when applied to large and complex models. Despite these challenges, Model-Agnostic Techniques remain invaluable tools for interpreting and explaining machine learning models in healthcare applications.

2.3 Counterfactual Explanations in Healthcare

By modeling alternative situations and analyzing their impact on patient outcomes, counterfactual explanation provides insightful information about cause-and-effect linkages in the healthcare industry.By contrasting actual patient outcomes with fictitious scenarios in which alternative treatments are used, it makes it possible to assess the efficacy of treatment. Healthcare practitioners can improve the accuracy of their predictive models by taking into account the possible impacts of certain patient characteristics or interventions on future outcomes by integrating counterfactual analysis. In order to improve patient care and enable healthcare professionals to make more educated treatment decisions, counterfactual explanations are vital decision support tools.

Various Counterfactual Explanations provide a range of methods for comprehending the causal connections found in machine learning models. WACH[5], for example, investigates theoretical situations by varying input factors and observing how they affect results, offering important information about the sensitivity of predictions. By creating

counterfactual scenarios that take sensitive characteristics into account, REVISED[6], on the other hand, gives ethical issues first priority and improves fairness and reliability in decision-making. Despite having different approaches, these strategies advance more egalitarian and transparent machine learning applications and deepen our understanding of model behavior.

Moreover, using a variety of counterfactual explanations, optimization-based methods such as DICE[7][8] seek to reflect the variability and uncertainty present in real-world circumstances. This technique offers thorough insights into possible outcomes connected to changes in input variables, which improves the interpretability and robustness of machine learning models. In a similar vein, LORE[9] focuses on producing instance-specific explanations that are customized for each unique situation, emphasizing customized comprehension and increasing the applicability of machine learning models.

Chapter 3

Research Gaps and Objectives

3.1 Resesarch Gaps

Many existing XAI techniques are computationally expensive and may not scale well to large datasets or complex models.While counterfactual explanations can provide insights into how changes in input features affect AI predictions, they do not always capture causal relationships.XAI aims to make AI systems more understandable to humans, there is limited research on how different stakeholders, such as end-users, domain experts, and policymakers, perceive and interact with XAI explanations.

Most XAI research assumes static datasets and models, but real-world environments are dynamic and evolving.The goal of XAI , counterfactual explanations and LIME methods is to make machine learning models more interpretable for humans, the explanations generated by these methods may still be difficult for non experts to understand. The process of generating perturbed samples and training surrogate models in LIME can be computationally intensive, especially for high-dimensional data or complex models.

3.2 Objectives

1. Use statistical techniques and machine learning on healthcare data to uncover causal connections.

2. Identify cause-and-effectrelationshipsin healthcare datasets while mitigating biases for fairness.

3. Employ various advanced XAI techniques like LIME, SHAP, and Integrated Gradients to explain causal relationshipstransparently.

4. Implement counterfactual explanationsto clarify data relationships, promoting transparency

5. Investigate the Impact of Fairness Constraints on Performance and Explainability in Deep Learning Models for Healthcare Data

Chapter 4

Proposed Methodology

In the field of healthcare, the suggested technique provides a thorough framework for creating and analyzing machine learning models. Finding appropriate benchmarking datasets is the first step, and then careful data pre-processing is done to guarantee the accuracy and consistency of the data. Next, in order to extract pertinent data and improve the prediction power of the model, feature engineering and selection are carried out. The performance of a variety of machine learning methods with the healthcare datasets is then used to evaluate and select them, including logistic regression, decision trees, and XGBoost.

Explainable AI (XAI) methods like SHAP, LIME, and partial dependency plots are used to analyze and interpret the model's predictions after it has been trained and assessed. These methods promote transparency by providing insights into the underlying causes influencing the model's judgments and trust among stakeholders. Moreover, fairness assessment is conducted to ensure equitable outcomes across different subgroups of the population. Through continuous monitoring and maintenance, the deployed model remains adaptive and responsive to evolving healthcare needs, thereby contributing to improved patient care and decision-making processes.

4.1 Process Workflow

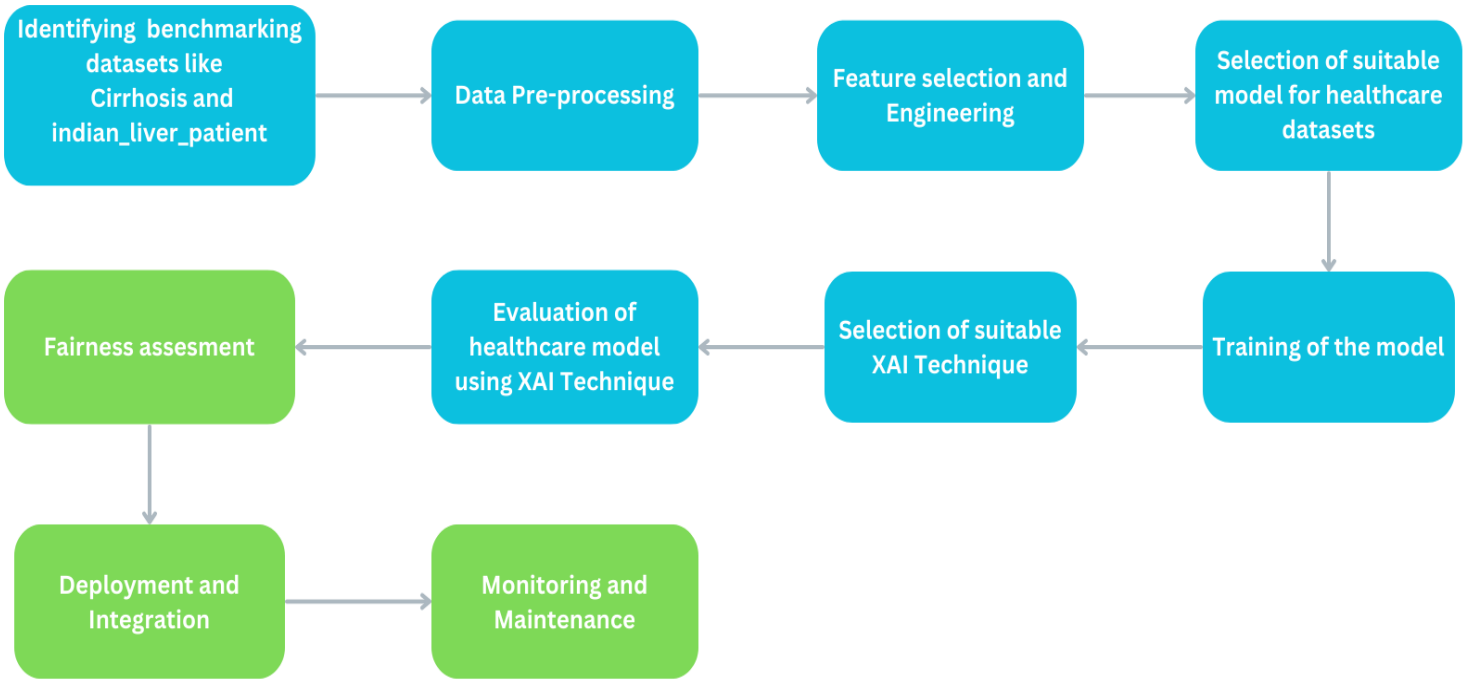

Figure 4.1: Process Workflow Diagram

The workflow consists of several steps, including:

1. Identifying Benchmarking Datasets: The first step in the process is to identify appropriate datasets to use in the project. This include datasets like Cirrhosis and indian liver patient. These datasets are used to train and evaluate the machine learning model.

2. Data Pre-processing: Once the datasets have been identified, then we pre-process the data. This involves cleaning the data, handling missing values, and transforming the data into a format that can be used by the machine learning model.

3. Feature Selection and Engineering: After the data has been pre-processed, the next step is to select and engineer the features that will be used in the machine

learning model. This involves identifying the most relevant features in the dataset and transforming them to improve the performance of the model.

4. Selection of Suitable Model for Healthcare Datasets: Once the features have been selected and engineered, the next step is to select an appropriate machine learning models for the healthcare dataset.

5. Training of the Model: After the machine learning model has been selected, the next step is to evaluate its performance. This involves training the model first then testing the model on a separate training set which is taken from the original datset itself and comparing its performance to other models.

6. Selection of Suitable XAI Technique: Once the model has been trained, then we select suitable XAI techniques like SHAP, LIME, Breakdown to generate explanations of the model.

7. Evaluation of healthcare model using XAI Technique: After the XAI model has been selected, the next step is to apply explainable artificial intelligence (XAI) techniques to help interpret the model's predictions. This involves identifying the most important features in the model and explaining how they contribute to the model's predictions.

8. Fairness Assessment using XAI Technique: Once the model has been trained and interpreted using XAI techniques, the next step is to assess its fairness. This involves evaluating the model's performance across different subgroups of the population and ensuring that it is not biased or discriminatory.

9. Deployment and Integration: After the model has been trained, interpreted, and evaluated for fairness, the next step is to deploy it in a real-world setting. This

involves integrating the model into a larger system or application, and ensuring that it is able to handle real-world data and use cases.

10. Monitoring and Maintenance: Finally, the last step in the workflow is to monitor and maintain the model over time. This involves tracking its performance, identifying and addressing any issues or biases that may arise, and ensuring that it remains up-to-date with the latest data and use cases.

4.2 Datasets

Here we have used two datasets:

1. Indian Liver Dataset
2. Cirrhosis

4.2.1 Indian Liver Dataset

This data collection was gathered from the northeast region of Andhra Pradesh, India, and includes 416 records of liver patients and 167 records of non-liver patients. Groups are classified as either liver patients (liver disease) or not (no disease) based on the class label "Dataset" in the column. There are 142 patient records for women and 441 patient records for men in this data collection.

Patients who are older than 89 are identified as being "90" years old.

Columns:

- Age of the patient
- Gender of the patient
- Total Bilirubin

- Direct Bilirubin
- Alkaline Phosphotase
- Alamine Aminotransferase
- Aspartate Aminotransferase
- Total Protiens
- Albumin
- Albumin and Globulin Ratio
- Dataset: field used to divide the data into two sets (patients with liver disease, or not)

4.2.2 Cirrhosis

The information gathered from the 1974–1984 Mayo Clinic trial on primary biliary cirrhosis (PBC) of the liver is presented here. During the course of those 10 years, 424 PBC patients were sent to Mayo Clinic and were found to be eligible for the D-penicillamine randomized placebo-controlled experiment. The first 312 instances in the dataset have data that is mostly full and were part of the randomized study. The extra 112 cases gave their permission to be monitored for survival and to have their baseline measurements taken, even if they chose not to take part in the clinical experiment. The results presented here include 312 randomized participants as well as an extra 106 cases, as six of those cases were lost to follow-up soon after diagnosis.

Attribute Information

- ID: unique identifier

- NDays: number of days between registration and the earlier of death, transplantation, or study analysis time in July 1986
- Status: status of the patient C (censored), CL (censored due to livertx), or D (death)
- Drug: type of drug D-penicillamine or placebo
- Age: age in [days]
- Sex: M (male) or F (female)
- Ascites: presence of ascites N (No) or Y (Yes)
- Hepatomegaly: presence of hepatomegaly N (No) or Y (Yes)
- Spiders: presence of spiders N (No) or Y (Yes)
- Edema: Y (edema in spite of diuretic medication), S (edema present without diuretics, or edema cleared by diuretics), or N (no edema and no diuretic therapy for edema)
- Bilirubin: serum bilirubin in [mg/dl]
- Cholesterol: serum cholesterol in [mg/dl]
- Albumin: albumin in [gm/dl]
- Copper: urine copper in [ug/day]
- AlkPhos: alkaline phosphatase in [U/liter]
- SGOT: SGOT in [U/ml]
- Triglycerides: triglicerides in [mg/dl]

- Platelets: platelets per cubic [ml/1000]
- Prothrombin: prothrombin time in seconds [s]
- Stage: histologic stage of disease (1, 2, 3, or 4)

4.3 Machine Learning Models

In this project we have used different models like logistic regression,KNN ,decision tree,XG Boost etc.. Here's a simple overview of each model:

1. Logistic Regression: A linear classification algorithm used for binary classification tasks, estimating the probability of a binary outcome based on input features.
2. K-Nearest Neighbors (KNN): A simple algorithm that classifies data points based on the majority class among their k-nearest neighbors in the feature space.
3. Decision Tree: A tree-like structure where each internal node represents a decision based on a feature, leading to splits that eventually classify data points into classes at the leaf nodes.
4. Support Vector Machine (SVM): A powerful classification algorithm that finds the optimal hyperplane to separate data points into different classes in a high-dimensional space.
5. Linear SVC: A variant of SVM for linear classification tasks, aiming to find the optimal hyperplane with a maximal margin between classes.
6. Random Forest: An ensemble learning method that builds multiple decision trees and combines their predictions to improve classification accuracy and robustness.

7. XGBoost: An optimized gradient boosting algorithm that sequentially builds a series of weak learners to minimize a differentiable loss function, achieving state-of-the-art results in many machine learning tasks.

8. Ridge Classifier: A linear classification algorithm that applies L2 regularization to penalize large coefficients, reducing model complexity and mitigating overfitting.

9. Dummy Classifier: A simple baseline classifier used for comparison, making predictions based on simple rules such as always predicting the majority class.

10. Linear Discriminant Analysis: A classification algorithm that projects data onto a lower-dimensional space, maximizing the separation between classes while minimizing intra-class variance.

4.4 Explainable AI Techniques

Here we have used different Explainable AI techniques like SHAP, LIME, Partial Dependence Plots and Breakdown

1. SHAP (SHapley Additive exPlanations): SHAP is a method used for explaining the output of machine learning models. It calculates each feature's contribution to the model's prediction for a given instance. Cooperative game theory's Shapley values serve as the foundation for SHAP values, which fairly distribute feature importance. Interpretability is improved with SHAP values, which provide information on how each feature affects the model's output. Neural networks, linear models, and tree-based models are just a few of the types to which SHAP can be applied.

2. LIME (Local Interpretable Model-agnostic Explanations): LIME is a method for providing local explanations for machine learning model predictions. It approx-

imates the behavior of the model around a particular instance to produce interpretable explanations. LIME generates an interpretable and easily comprehensible local surrogate model close to the instance of interest. Due to LIME's model-agnostic nature, any black-box model can be used without the need to understand its internal workings. LIME enables users to trust and comprehend complex models by offering insights into a model's behavior locally.

3. BreakDown: BreakDown is a method for breaking down a machine learning model's prediction into the contributions of individual features. It is simple to read and analyze since it provides a clear explanation of how each feature affects the final prediction. BreakDown sheds light on the significance of features and the directionality of their influences on the output of the model. It's especially helpful for providing clear, understandable explanations for individual predictions. BreakDown can be used to improve the transparency and reliability of a variety of models, such as neural networks, tree-based models, and linear models.

Chapter 5

Experimental Results

The project report's experimental results section provides a detailed analysis of machine learning models used with healthcare datasets, with a particular emphasis on liver disease prediction. The distribution and interactions between different characteristics are efficiently communicated through visualizations such as correlation matrices and histograms, which help with feature selection and preprocessing. Furthermore, comparing the accuracies of various algorithms offers valuable information about how well they predict liver illness, which helps with well-informed model selection for clinical applications.

Furthermore, the models' interpretability is improved by the incorporation of explainable AI (XAI) approaches such variable significance plots and SHAP values. These methods provide researchers and physicians with insightful information by clarifying the influence of particular traits on predictions, leading to a better comprehension of the fundamental elements controlling liver disease prediction.

5.1 Distributions Of Features

Below are the histograms of features for the two datasets. The above histogram plot

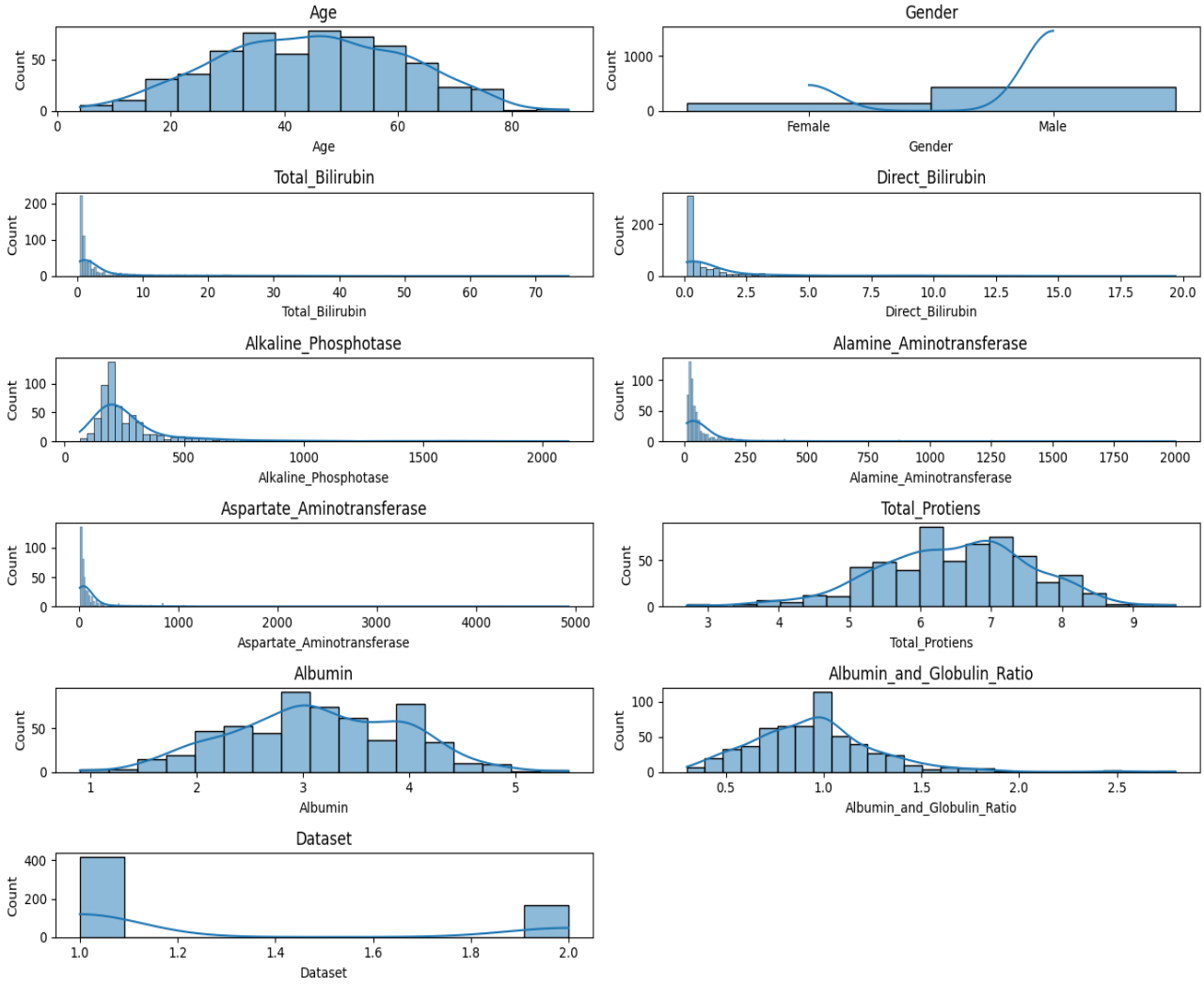

Figure 5.1: Histogram of Indian Liver Dataset

shows the distribution of various features in a healthcare dataset. The plot includes several histograms, each representing a different feature in the dataset. The features include demographic information (e.g., age, gender), lab test results (e.g., total bilirubin,

alkaline phosphotase, aspartate aminotransferase, albumin, direct bilirubin), and other relevant patient information (e.g., total proteins, albumin and globulin ratio).

Each histogram shows the frequency of the values for a particular feature, with the x-axis representing the value range and the y-axis representing the frequency. The plot includes several histograms for each feature, with different bins and color schemes to help distinguish between them.

For example, the first histogram shows the age distribution of the patients, with values ranging from 20 to 40 years. The second histogram shows the distribution of total bilirubin values, with values ranging from 0 to 30.

Overall, this histogram plot provides a useful summary of the distribution of various features in the healthcare dataset, which can help inform the selection and pre-processing of the data for machine learning models.

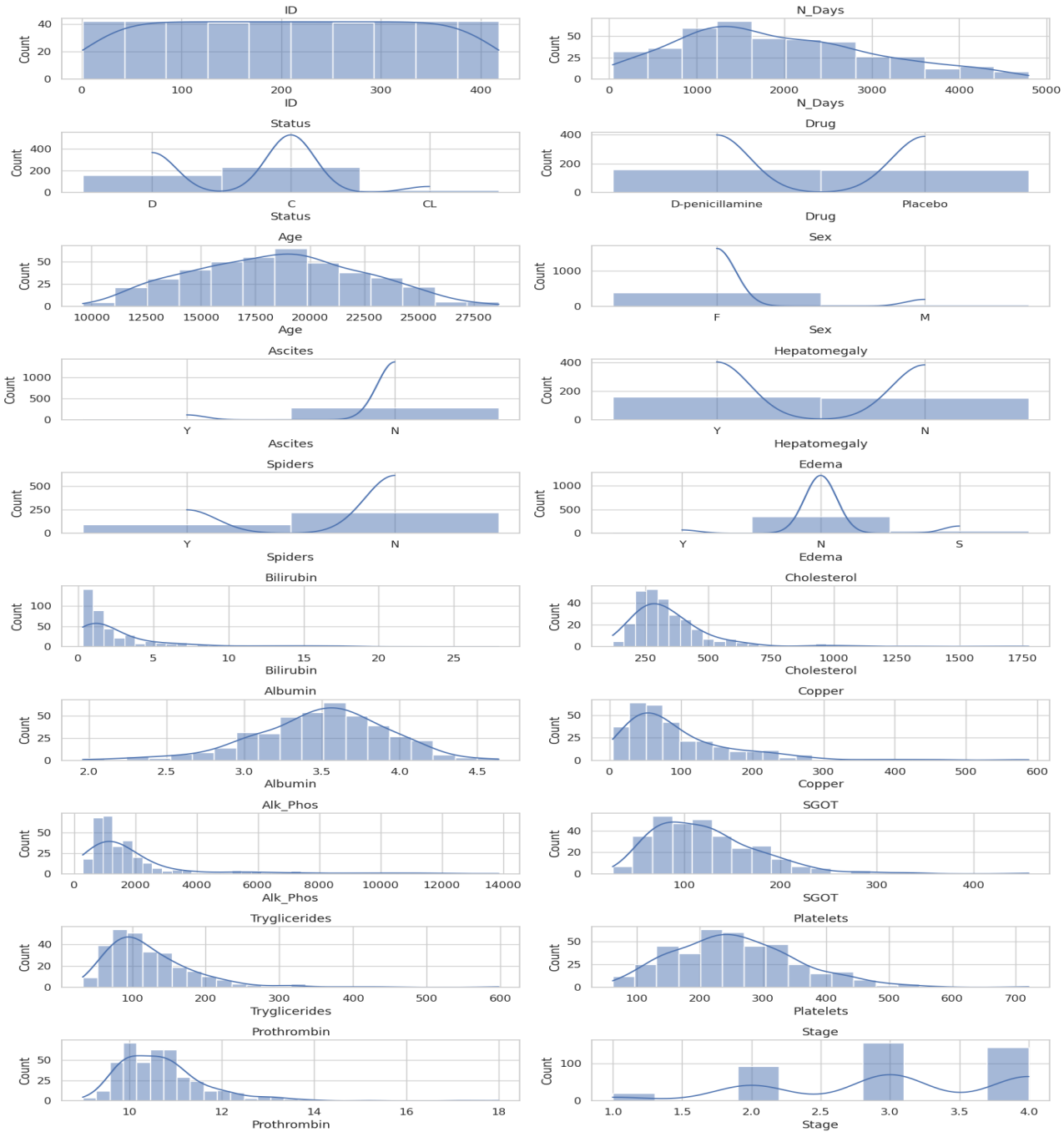

Figure 5.2: Histogram of Cirrhosis Dataset

The above histogram plot shows the distribution of various features in a healthcare dataset. Each histogram represents a different feature in the dataset, and the x-axis shows the value range for that feature. The y-axis shows the frequency of each value in the dataset.

The plot includes histograms for demographic information (such as age and sex), lab test results (such as bilirubin, albumin, and alkaline phosphatase), and other relevant patient information (such as ascites, hepatomegaly, and edema).

The plot shows that some features have a normal distribution (such as age and albumin), while others have a skewed distribution (such as bilirubin and alkaline phosphatase). Additionally, some features have distinct clusters (such as sex and ascites), while others have a more continuous distribution (such as albumin and bilirubin).

Overall, the histogram plot provides a useful summary of the distribution of various features in the healthcare dataset, which can help inform the selection and pre-processing of the data for machine learning models.

5.2 Correlation Matrix

The feature correlation matrices for the two datasets are shown below. A negative correlation is shown by a correlation coefficient between -1 and 0. Positive correlation is indicated by a correlation coefficient between 0 and 1. A perfect positive correlation is represented by a correlation coefficient of 1, and a perfect negative correlation is represented by a correlation coefficient of -1. There is no association when the correlation coefficient is 0.

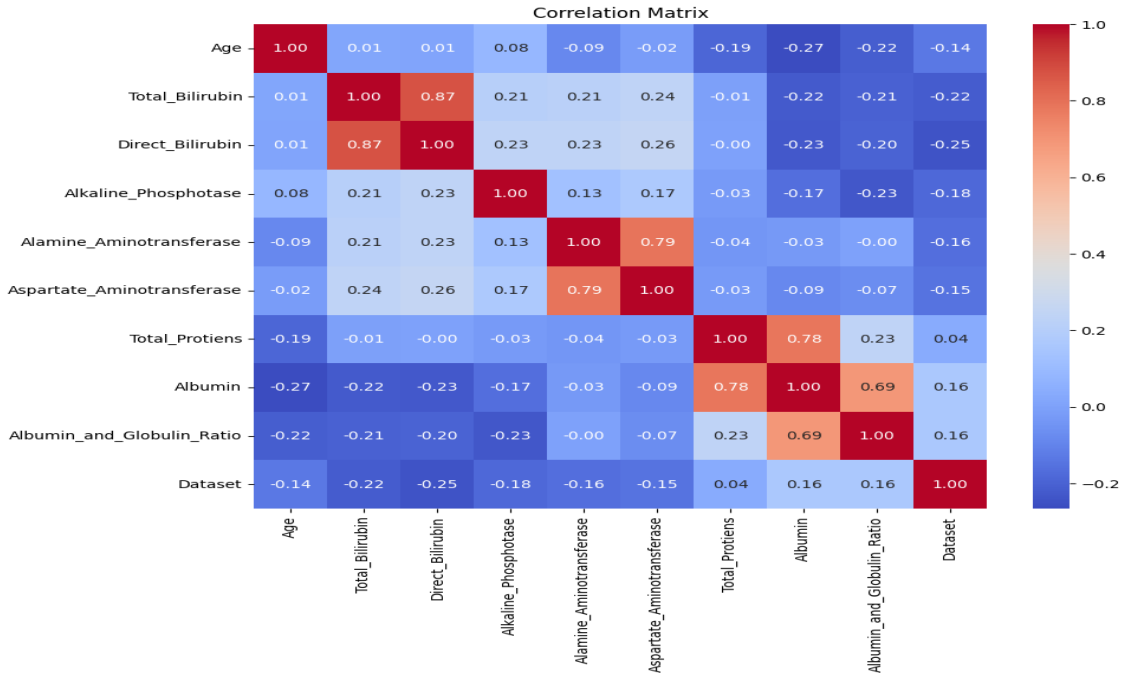

Figure 5.3: Correlation matrix of Indian Liver Dataset

Age and total bilirubin have a positive connection (0.01) in the plot. This indicates that total bilirubin tends to increase with age but in very less. Total bilirubin and direct bilirubin have a positive association (0.87), meaning that when total bilirubin rises, so does direct bilirubin.

Age and alkaline phosphatase have a marginally favorable connection (0.08) Age and albumin have a marginally negative connection (-0.27) Alkaline phosphatase and total bilirubin have a positive connection (0.21). Alkaline phosphatase and direct bilirubin have a positive connection (0.23) The relationship between aspartate aminotransferase and alanine aminotransferase is favorable (0.79) The relationship between total protein and albumin is positive (0.78).

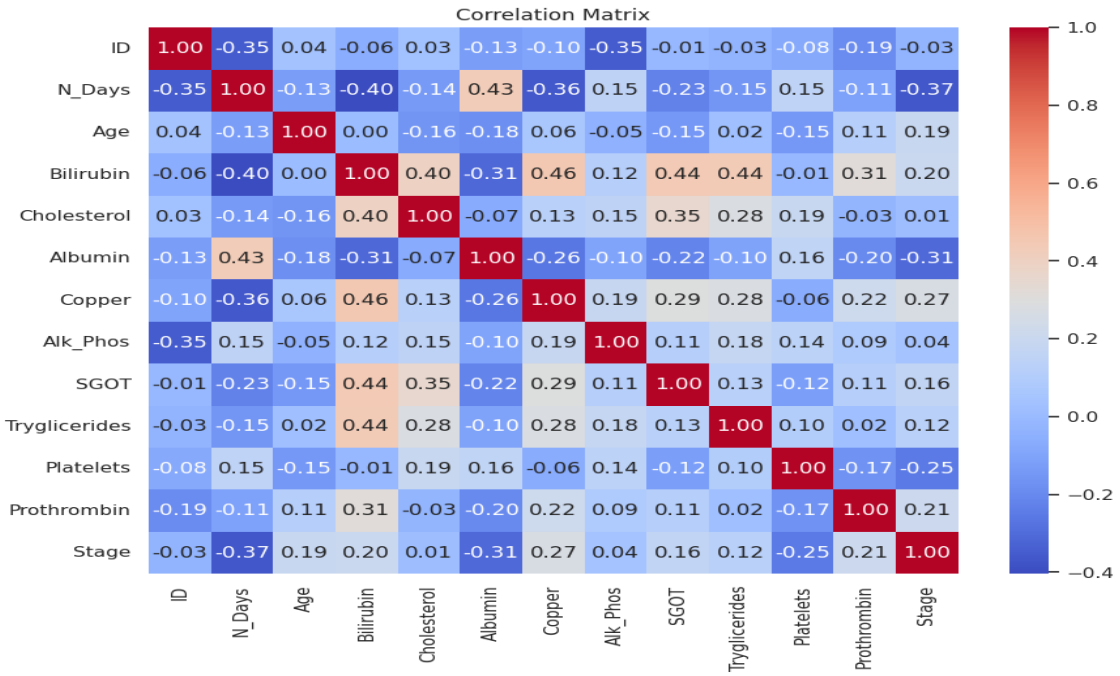

Figure 5.4: Correlation matrix of Cirrhosis Dataset

Age and albumin have a weakly negative connection (-0.18) in the plot. This indicates that albumin levels tend to decline with aging. Bilirubin and cholesterol have a positive association (0.40), meaning that when bilirubin levels rise, cholesterol levels likewise likely to rise.

Age and alkaline phosphatase have a marginally negative connection (-0.05). Alkaline phosphatase and bilirubin have a positive connection (0.12). The relationship between bilrubin and copper is favorable (0.46). The relationship between bilrubin and tryglicerides is positive (0.44). The relationship between albumin and platelets is marginally positive (0.16).

5.3 Accuracy Comparision

Below is The table representing the accuracy for both the datasets on different models.

Model Name	Indian Liver Patient	Cirrhosis
Logistic Regression	**76**	**78**
K-Nearest Neighbours	75	72
Decision Tree	70	68
Support Vector Machine	74	77
Linear SVC	76	77
Random Forest	73	75
XG Boost	75	74
Ridge Classifier	74	76
Dummy Classifier	74	52
Linear Discriminant Analysis	75	78

Table 5.1: Accuracy of Different Models

The table provided shows the accuracy of various machine learning models when applied to two healthcare datasets, namely the Indian Liver Patient dataset and the Cirrhosis dataset. The table consists of 3 columns and 10 rows, with the first column listing the names of the models, and the remaining 2 columns displaying the accuracy of each model for the two datasets.

According to the findings, the majority of the models functioned really well, with several of them reaching accuracy levels of 70% or higher. With an accuracy of 76% for the Indian Liver Patient dataset and 78% for the Cirrhosis dataset, the Logistic Regression model demonstrated the highest level of performance. It is crucial to remember that any model's performance can differ based on the particular dataset and features employed.

5.4 Explainable AI Techniques

5.4.1 Variable Importance

In eXplainable Artificial Intelligence (XAI), variable importance charts are frequently employed to offer insights into the relative relevance of characteristics or variables utilized in prediction models. The ability to comprehend how various input features affect model predictions is made possible by these charts, which improves transparency and interpretability—two things that are crucial for many real-world applications.

Below are the variable Importance plots on the two datasets.

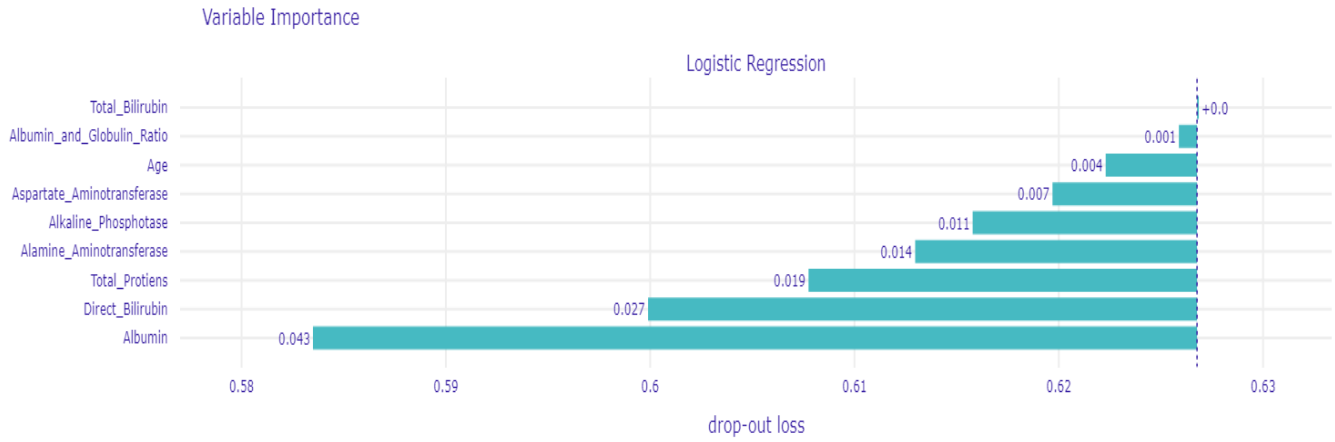

Figure 5.5: Variable Importance of Indian Liver Dataset

The relative significance of each variable in a logistic regression model for predicting liver disease in patients is represented graphically in the variable importance plot seen above. Based on the Indian Liver Patient dataset, which contains a range of laboratory and clinical variables, the figure is produced.

Total bilirubin is the most significant variable, followed by age and albumin and globulin ratio in that order of importance. The significance of each variable is reflected in the length of each bar in the plot; longer bars denote more significant predictors of

liver disease.

The high importance of Age suggests that age-related factors may play a significant role in the development of liver disease, and further research is needed to understand the specific mechanisms involved. Overall, the variable importance plot is a valuable tool for identifying the most important predictors of liver disease and can help inform the development of more effective diagnostic and treatment strategies in the future.

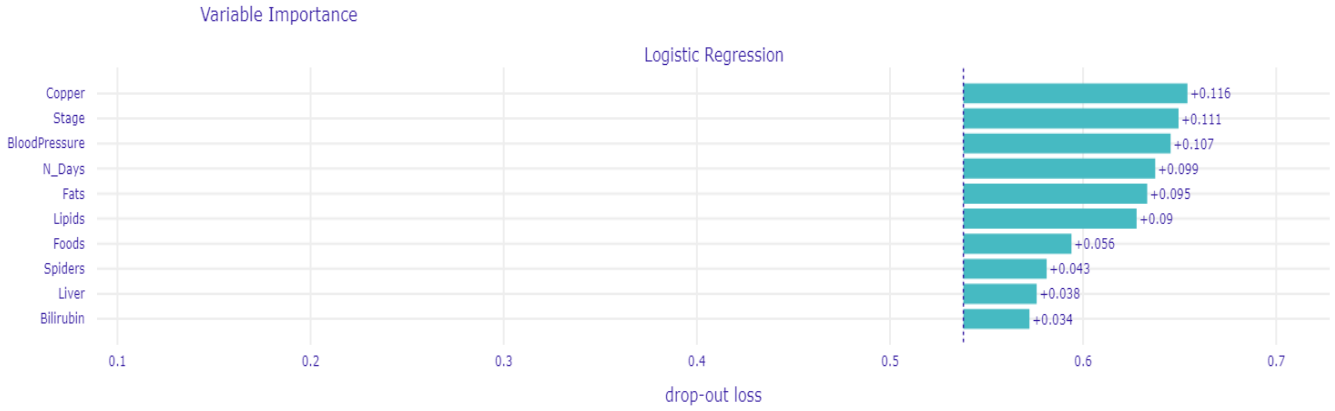

Figure 5.6: Variable Importance of Cirrhosis Dataset

Copper is the most significant variable, followed by Stage, Blood Pressure, N_Days, Fats, Lipids, Foods, Spiders, Liver, and Bilirubin. The factors are sorted in order of significance. The significance of each variable is reflected in the length of each bar in the plot; longer bars denote more significant predictors of liver disease.

According to Stage, the second most significant variable, the severity of liver disease may also be a significant predictor of the course of liver disease. Similar to blood pressure, lipids, and fats, these factors have a modest impact, indicating that they might have a role in the onset and progression of liver disease.

The remaining factors, on the other hand, such as bilirubin, food, spiders, liver, and N_Days, are not as significant. This suggests that these variables may have limited utility

in predicting liver disease, and further investigation may be needed to determine their relevance in the diagnostic and treatment pathway.

5.4.2 SHAP on whole Model

SHAP (SHapley Additive exPlanations) values offer profound insights into the inner workings of machine learning models, particularly in understanding feature importance and the impact of individual features on model predictions. When applied to the entire dataset, SHAP values provide a comprehensive overview of how each feature contributes to the model's predictions across the entire data distribution. Below are the plots which describes the shap values of different features and their imact on the model output.

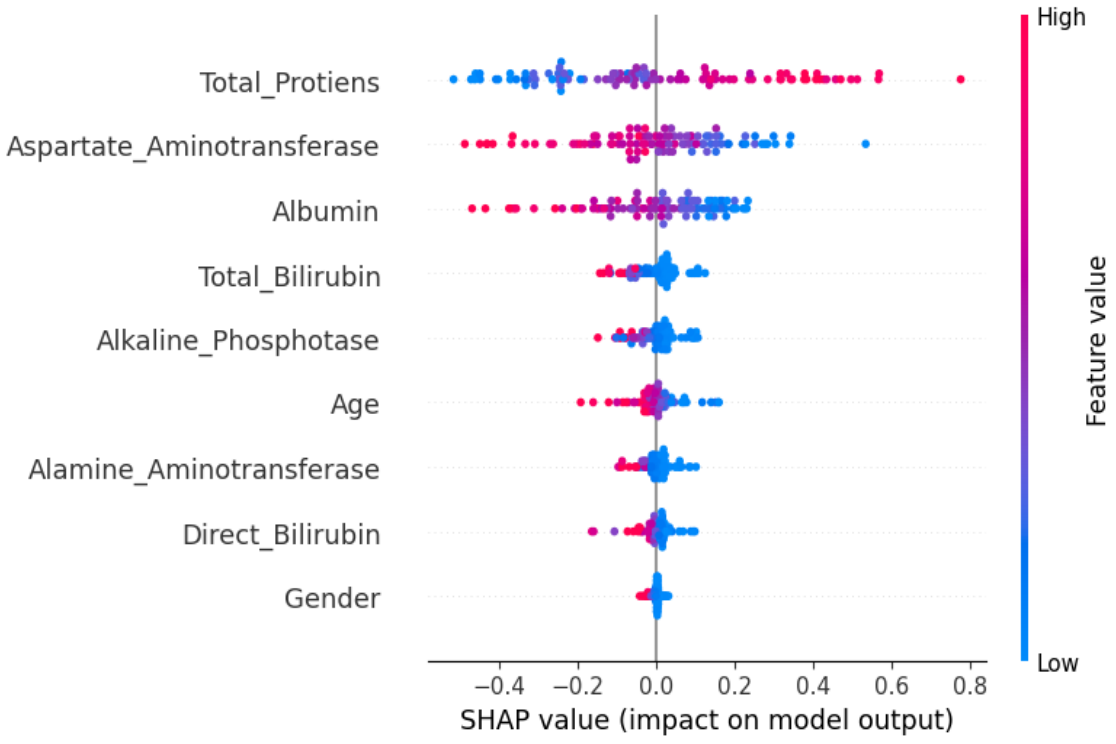

Figure 5.7: SHAP on Indian Liver Dataset

The feature significance plot produced by a logistic regression model trained on the In-

dian liver disease dataset is examined in this plot. The most important variables that the model took into account when determining whether liver disease will be present or absent are depicted in the plot. Plotting the data allows us to see which features—Total_Protiens and Asparte_Aminotransferase, for example—have the greatest bars, indicating that they significantly influence the model's predictions. These characteristics most likely indicate the most useful clinical markers of liver disease in the dataset.

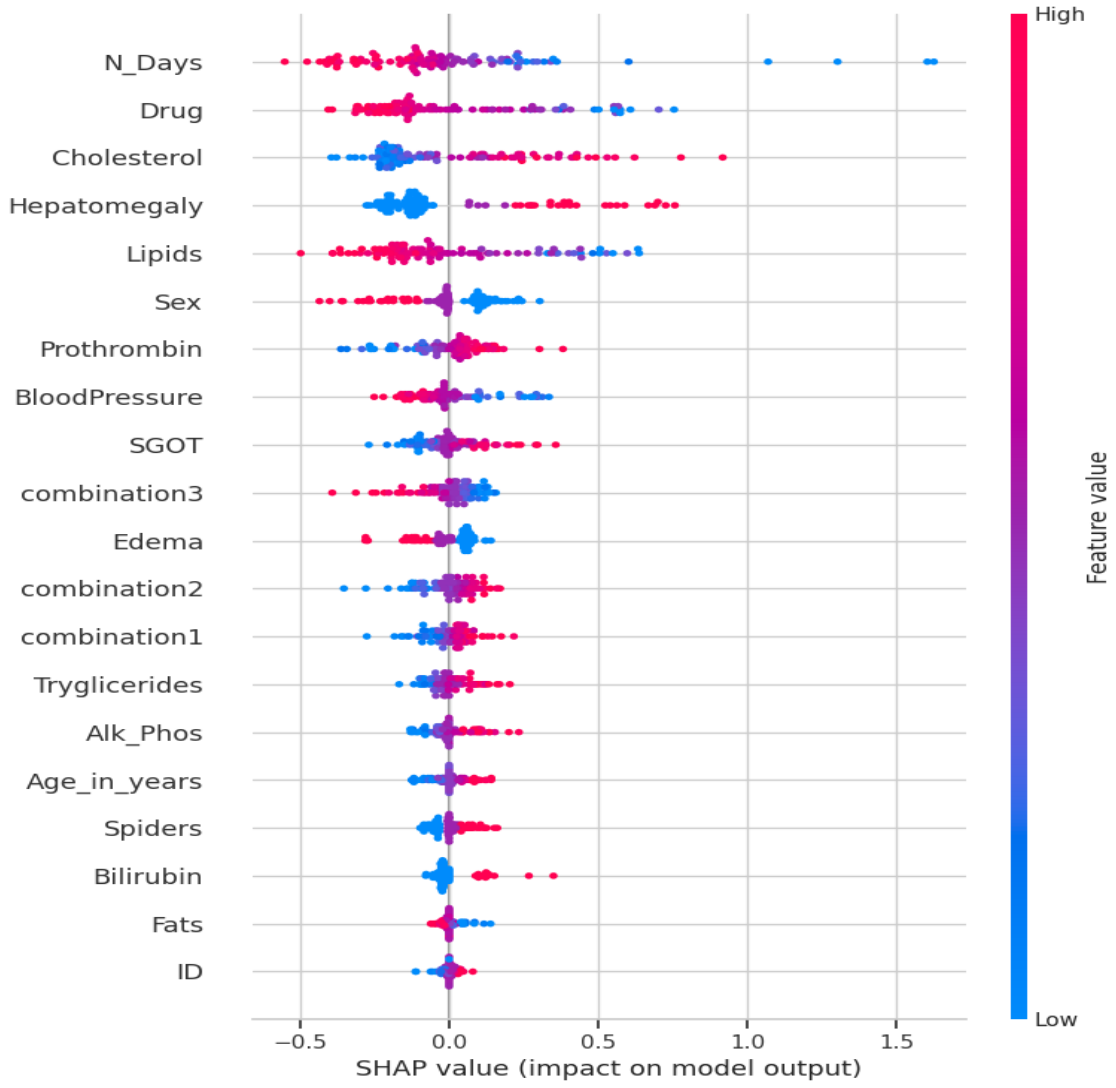

Figure 5.8: SHAP on Cirrhosis Dataset

The feature significance plot produced by a logistic regression model trained on the cirrohsis is analyzed in this plot. The most important variables that the model took into account when determining whether liver disease will be present or absent are depicted in the plot. Plot analysis allows us to determine which features—N_Days, Drug, Cholestrol, and Lipids—have the greatest bars, indicating that they significantly influence the model's predictions. These characteristics most likely indicate the most useful clinical markers of liver disease in the dataset.

5.4.3 SHAP vs LIME vs Breakdown

SHAP plots explain how each feature in a dataset contributes to a single prediction made by a machine learning model.SHAP plot visually represents how each feature (e.g., bilirubin levels, enzyme activity levels) in a specific patient's data point influences the model's prediction of whether that patient has liver disease or not.

LIME (Local Interpretable Model-agnostic Explanations) values offer another perspective on model interpretability, particularly focusing on providing explanations for individual predictions rather than the global behavior of the model.

Both LIME and SHAP provide explanations for individual predictions, but they use different approaches. LIME builds a simpler model around the data point, while SHAP utilizes game theory to attribute prediction weights to features.

Breakdown plots are a valuable tool for understanding how individual features contribute to model predictions across the entire dataset. These plots provide a clear visual representation of feature effects, allowing for a nuanced analysis of how changes in each feature influence the model's output. This plot explains how the model arrived at a prediction for this particular patient by showing the contribution of each feature to the final prediction.

Below are the plots comparing SHAP , LIME and breakdown for a sample of each dataset.

Indian Liver Patient Dataset

Below plots correspond to a sample in Indian Liver Dataset

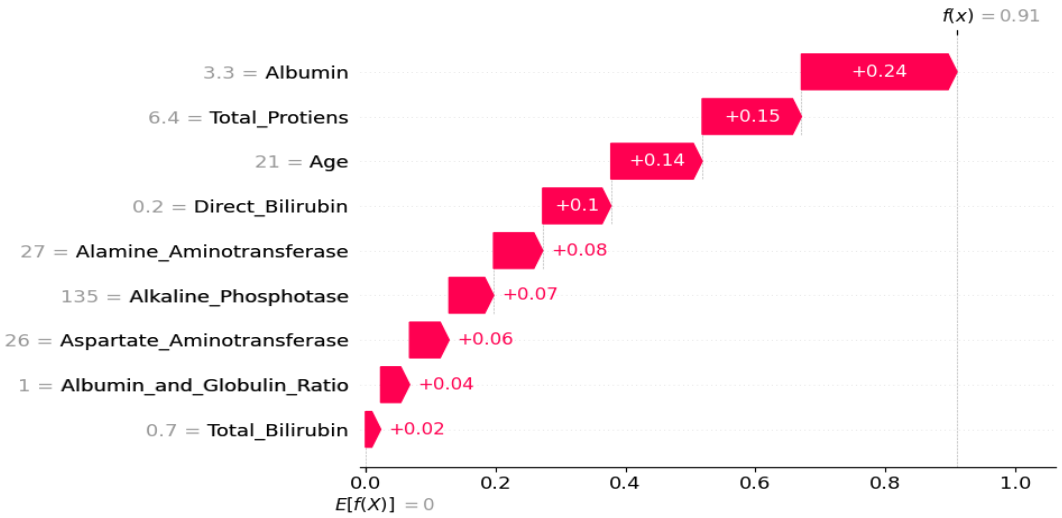

Figure 5.9: SHAP on a sample in Indian Liver Dataset

The horizontal dashed line in the middle represents the average prediction (baseline) of the model for liver disease across all data points. The location of each feature's bar relative to this line indicates how much that feature's value for this particular patient shifted the prediction from the baseline.

The positive SHAP value for Albumin (around 0.24) suggests a higher albumin level than usual, which pushes the prediction away from liver disease (since albumin is often lower in liver disease cases).The positive SHAP value for Age (around 0.14) indicates an older patient's age contributes a small nudge towards predicting liver disease.

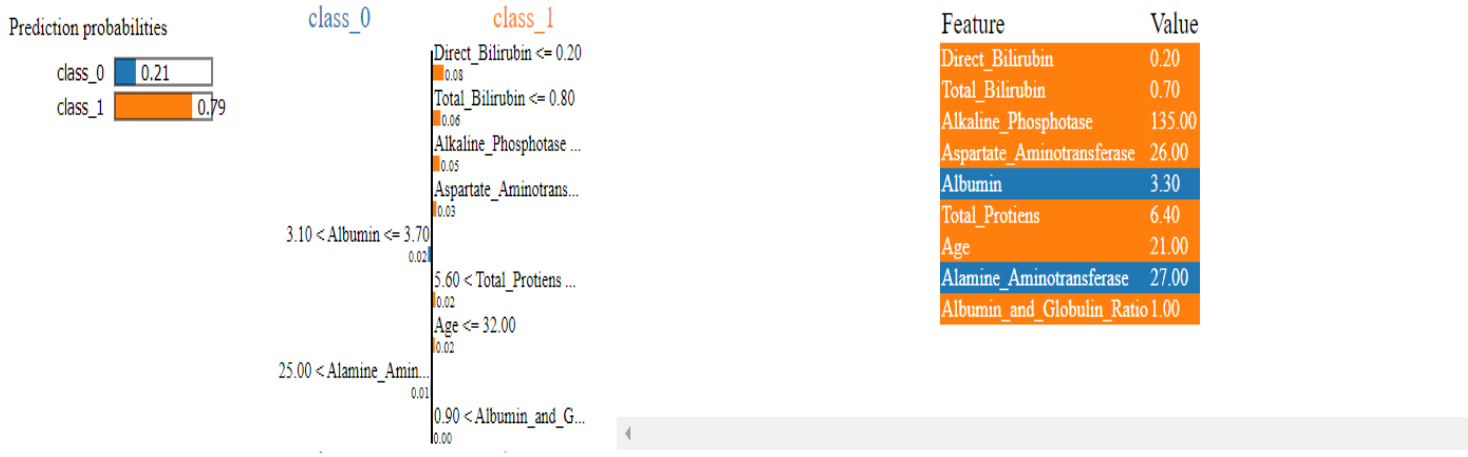

Figure 5.10: LIME on a sample in Indian Liver Dataset

The prediction probabilities are given for two classes: class_0 and class_1. The probability of the sample belonging to class_0 is 0.21, and the probability of the sample belonging to class_1 is 0.79.The probability of the sample belonging to class_0 if the value of the "Albumin" feature is between 3.10 and 3.70 is 0.02. Similarly, the probability of the sample belonging to class_1 if the value of the "Direct Bilirubin" feature is less than or equal to 0.20 is 0.08.

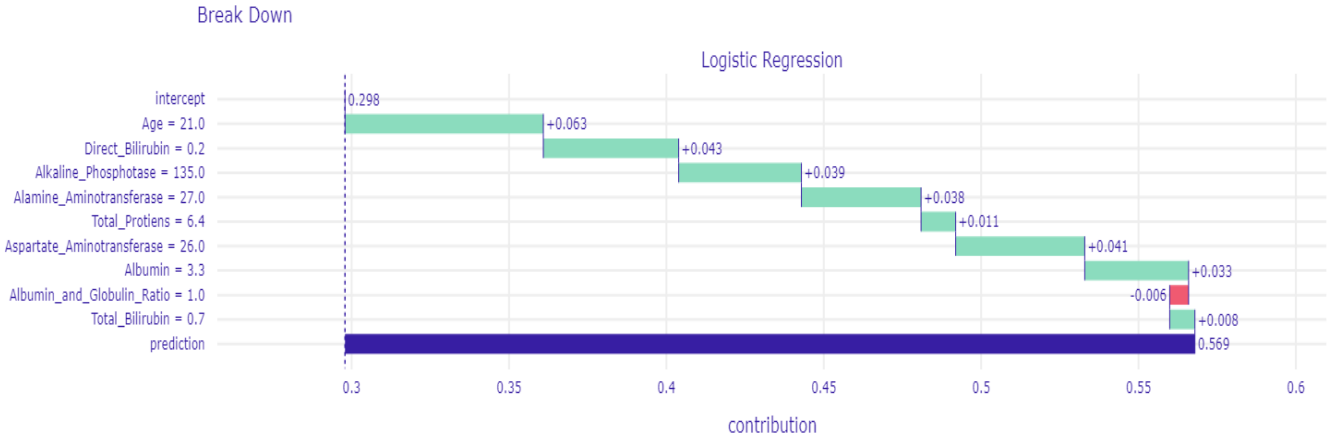

Figure 5.11: Breakdown on a sample in Indian Liver Dataset

This intercept value represents the baseline odds of a patient having liver disease according to the model, independent of any specific features. The value next to the feature name (e.g., 0.063 for Age) indicates the contribution of that specific feature value for this patient to the model's prediction. A positive value signifies the feature value increased the model's prediction of liver disease for that patient. Conversely, a negative value indicates the feature value decreased the prediction of liver disease.

Cirrhosis Prediction Dataset

Below plots correspond to a sample in Cirrhosis dataset.

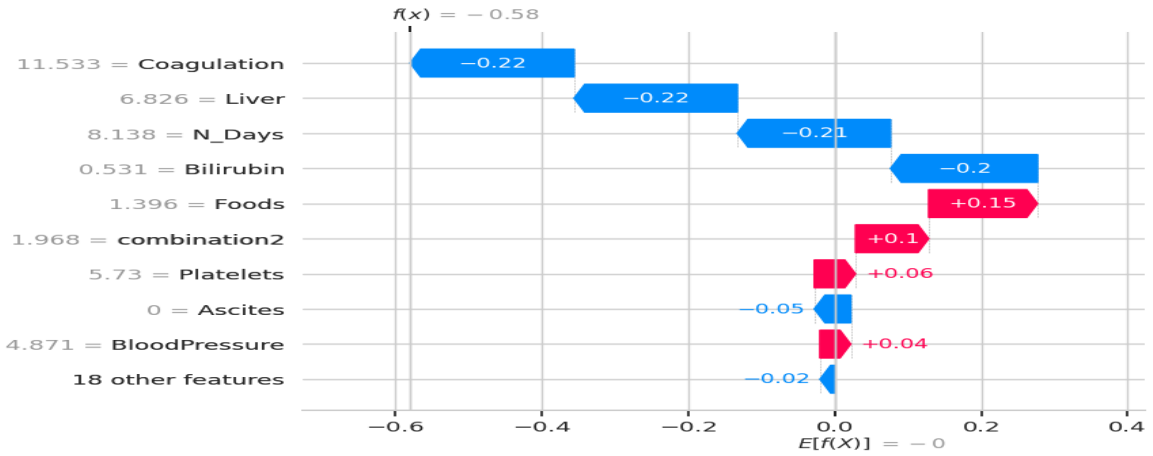

Figure 5.12: SHAP on a sample in Cirrhosis Dataset

Features with Positive SHAP Values indicates These features contribute to the model predicting a higher likelihood of liver disease for that patient. Examples might include elevated foods or blood pressure. Features with Negative contribute to the model predicting a lower likelihood of liver dataset for that patient. Examples might include cogulation.

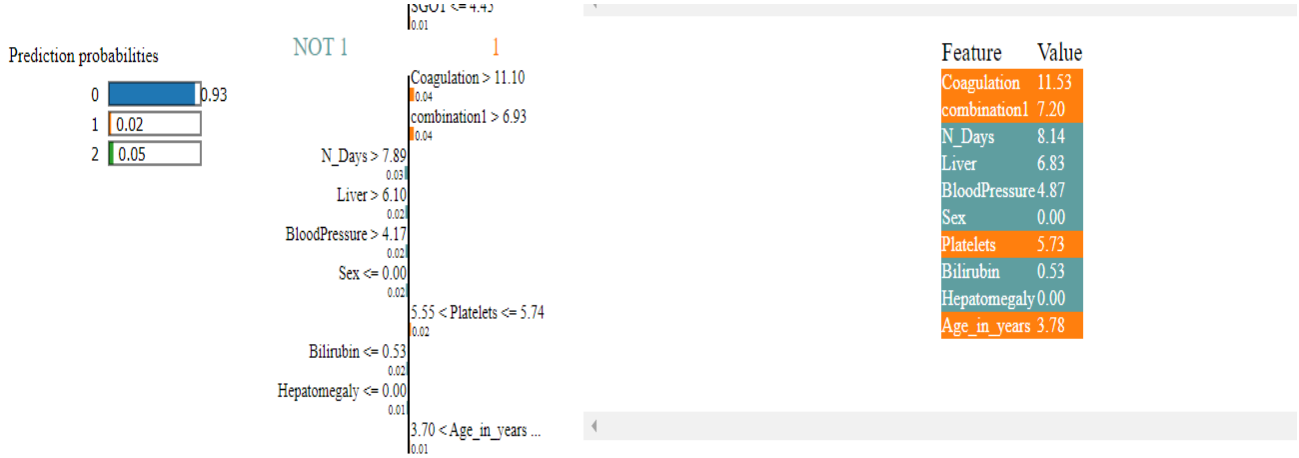

Figure 5.13: LIME on a sample in Cirrhosis Dataset

The following rows of the table show the prediction probabilities for certain ranges of values for various features in the dataset. For example, the probability of the sample belonging to cirrhosis (class 1) if the value of the "N_Days" feature is between 7.89 and 8.14 is 0.03. Similarly, the probability of the sample belonging to cirrhosis (class 1) if the value of the "Liver" feature is between 6.10 and 6.83 is 0.02.

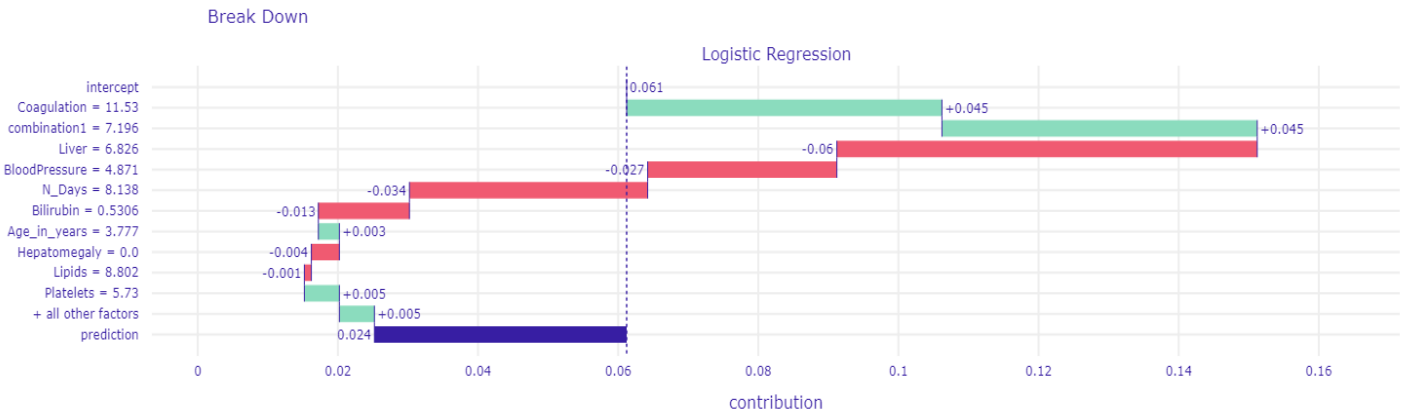

Figure 5.14: Breakdown on a sample in Cirrhosis Dataset

The plot shows the intercept for the Coagulation feature, which is 11.53. This means

that the intercept has the most significant contribution to the prediction, resulting in a probability of 0.16.Moving to the right, the plot shows the contribution of the combination1 feature, with a value of 7.196. This value is associated with a negative contribution to the prediction, which reduces the probability to 0.14.

The next factors in the plot are the Liver and Blood Pressure features, with values of 6.826 and 4.871, respectively. Both of these values are associated with a positive contribution to the prediction, which increases the probability to 0.08 and 0.04, respectively. The N_Days and Bilirubin features have negative contributions to the prediction, reducing the probability to 0.02 and 0.06, respectively. The Age_in_years and Lipids features both have positive contributions to the prediction, but the contribution is relatively small, resulting in minor changes to the probability.

5.5 Summary of Explainable AI Techniques

In summary below are the comparison plots of SHAP, LIME, Breakdown plots and there variable importance. Here length bar represents the variable importance of that particular feature generated by that particular XAI Technique. Each XAI Technique is represented by different bar type.

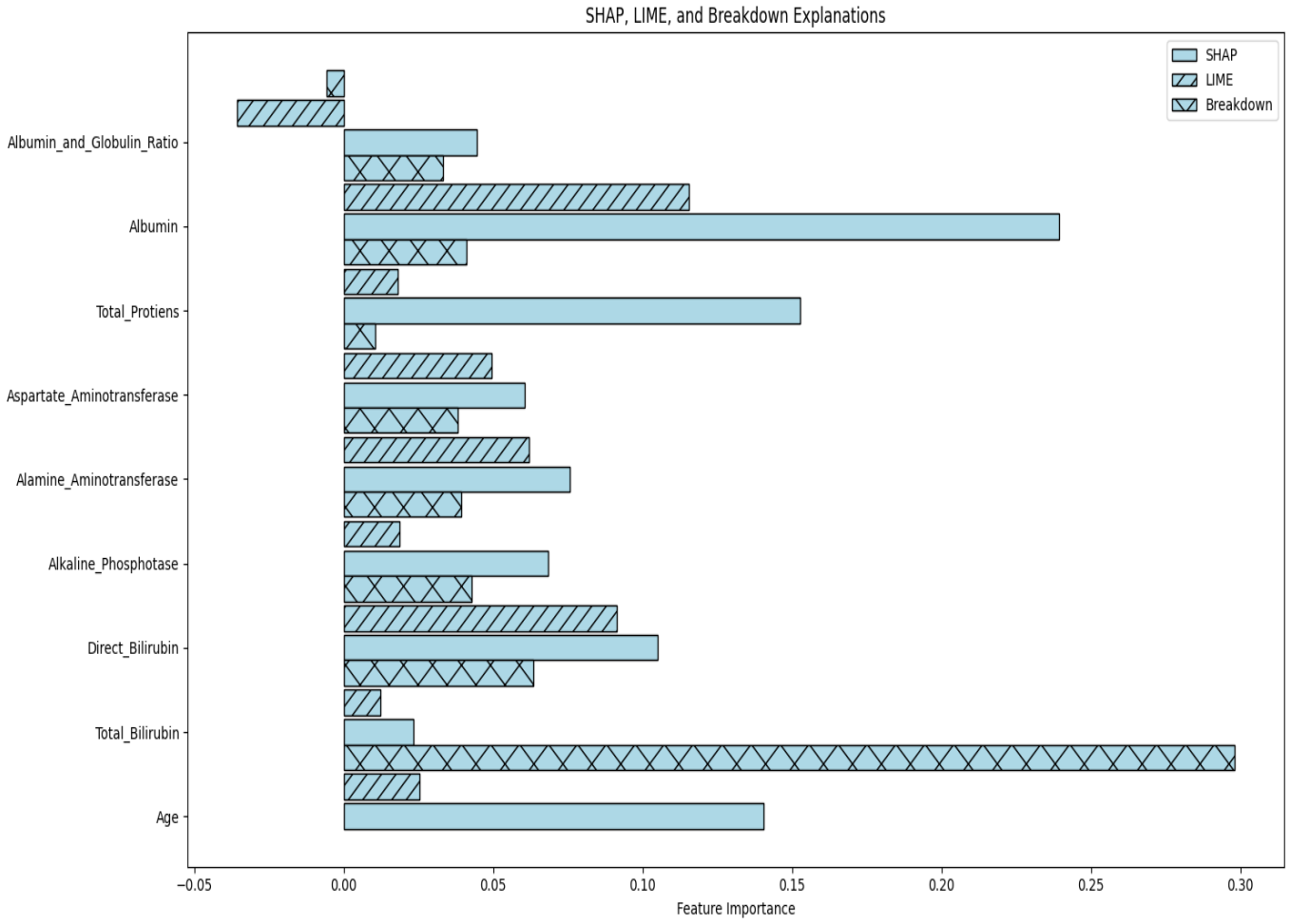

Figure 5.15: Summary on a sample in Indian Liver Patient Dataset

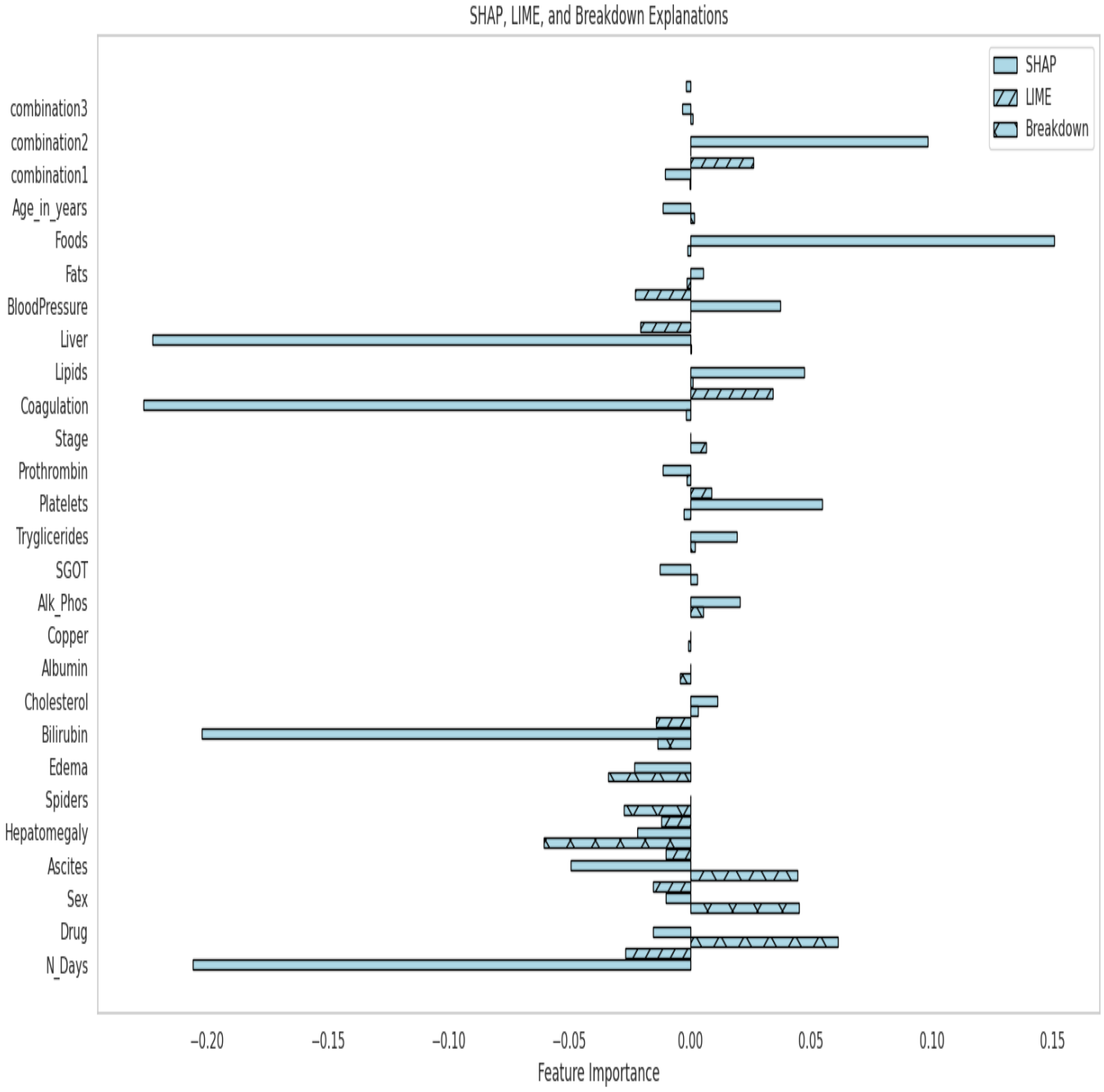

Figure 5.16: Breakdown on a sample in Cirrhosis Dataset

From the above plots we can see the the variable importance value of each feature in that particular dataset represented by the bars side by side for each XAI technique.

In liver prediction dataset using SHAP Albumin, Total_Protiens, Age and Direct_Bilirubin comes out to be the most important features. While using LIME on that particular sample itself Albumin, Direct_Bilirubin, Aspartate_Aminotransferase, Alamine_Aminotransferase comes out to be the most important features. While using Breakdown Age, Total_Bilirubin, Direct_Bilirubin and Alkaline_Phosphate comes out to be the most important features.

In Cirrhosis dataset using SHAP Liver, N_Days, Coagulation and Bilirubin comes out to be the most important features. While using LIME on that particular sample itself N_Days, Coagulation, Combination1 and Liver comes out to be the most important features. While using Breakdown N_Days, Hepatomegaly, Drug and Sex comes out to be the most important features.

Chapter 6

Conclusion

We utilized two distinct datasets, namely the Indian Liver Patient dataset and the cirrhosis dataset. By employing a diverse array of machine learning algorithms including Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Linear SVM, Decision Trees, XGBoost, and Random Forest, we aimed to comprehensively explore the predictive capabilities of these models in the context of liver disease diagnosis.

Through extensive experimentation and evaluation, we meticulously assessed the performance of each model by considering metrics such as accuracy, precision, recall, and F1-score. This comprehensive evaluation provided valuable insights into the strengths and weaknesses of each algorithm, allowing us to identify the most effective approach for liver disease prediction.

Our findings revealed that while certain models exhibited high accuracy rates, others demonstrated superior performance in terms of sensitivity or specificity. This highlights the importance of considering multiple evaluation metrics to gain a holistic understanding of model performance.

Moreover, our exploration of eXplainable Artificial Intelligence (XAI) techniques, particularly SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-

agnostic Explanations), and Breakdown plots, further enriched our analysis by providing interpretable insights into the decision-making process of the Logistic Regression model.

By leveraging these XAI techniques, we were able to elucidate the underlying factors driving the model's predictions, thereby enhancing transparency and interpretability. This not only facilitated a deeper understanding of the features contributing to the risk of liver disease but also empowered medical practitioners with valuable insights that could inform clinical decision-making.

Furthermore, our project underscored the significance of explainability in machine learning models, particularly in healthcare applications where interpretability is paramount. By elucidating the rationale behind the model's predictions, we fostered trust and confidence in the predictive capabilities of our approach, ultimately advancing the field of liver disease diagnosis and treatment.

In conclusion, our project not only contributed to the development of accurate and interpretable models for liver disease prediction but also underscored the importance of transparency and explainability in healthcare-related machine learning applications.

Chapter 7

Future Workplan

In the upcoming phases we would like to build different Deep learning models and compare their accuracy on both the datasets. We would also like to try and compare different XAI techniques on the best deep learning model obtained. We would also extend our project for trying other XAI techniques other than SHAP, LIME and Breakdown. Going further we would like to implement fairness assesment on the models we built.

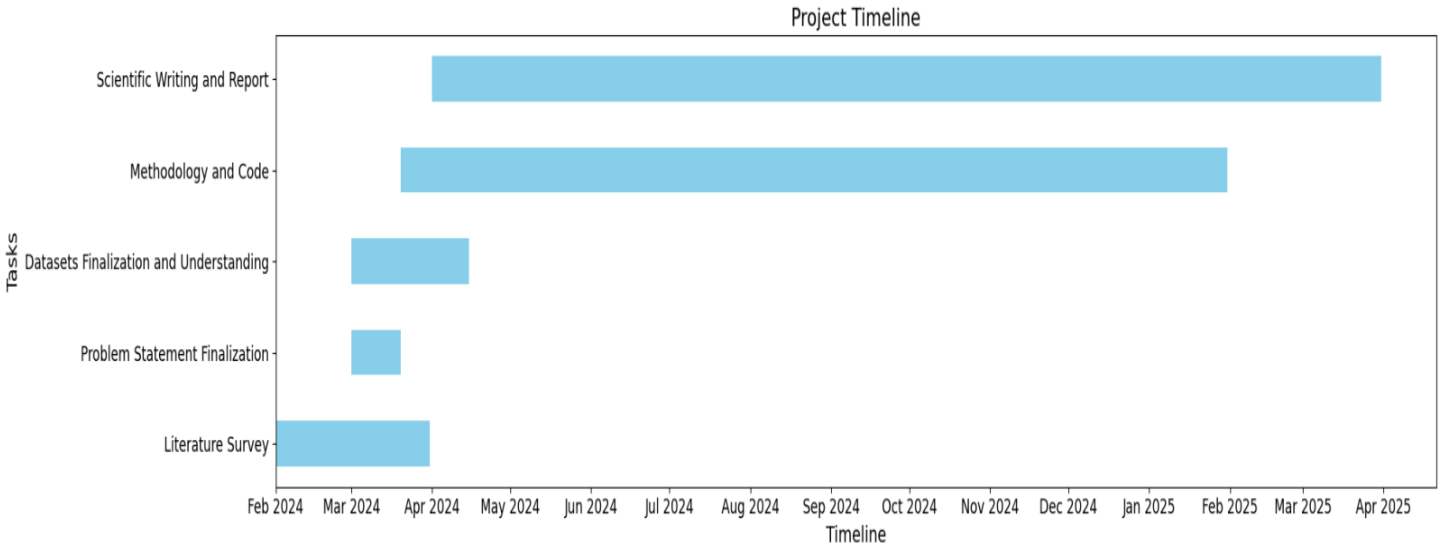

Figure 7.1: Future Workplan

References

[1] Devam Dave, Het Naik, Smiti Singhal, and Pankesh Patel. Explainable ai meets healthcare: A study on heart disease dataset. 2020.

[2] Yiming Zhang, Ying Weng, and Jonathan Lund. Applications of explainable artificial intelligence in diagnosis and surgery. *Diagnostics*, 12:237, 01 2022.

[3] Jörn Lötsch, Dario Kringel, and Alfred Ultsch. Explainable artificial intelligence (xai) in biomedicine: Making ai decisions trustworthy for physicians and patients. *BioMedInformatics*, 2, 12 2021.

[4] Krishnaraj Chadaga, Srikanth Prabhu, Niranjana Sampathila, Rajagopala Chadaga, Shashikiran Umakanth, Devadas Bhat, and Shashi S. Explainable artificial intelligence approaches for covid-19 prognosis prediction using clinical markers. *Scientific Reports*, 14:1783, 01 2024.

[5] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 607–617, New York, NY, USA, 2020. Association for Computing Machinery.

[6] Javier Del Ser, Alejandro Barredo-Arrieta, Natalia Díaz-Rodríguez, Francisco Herrera, Anna Saranti, and Andreas Holzinger. On generating trustworthy counterfactual explanations. *Information Sciences*, 655:119898, 2024.

[7] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *CoRR*, abs/1711.00399, 2017.

[8] Ramaravind Kommiya Mothilal, Divyat Mahajan, Chenhao Tan, and Amit Sharma. Towards unifying feature attribution and counterfactual explanations: Different means to the same end. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 652–663, New York, NY, USA, 2021. Association for Computing Machinery.

[9] Riccardo Guidotti, Anna Monreale, Fosca Giannotti, Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems*, 34:14–23, 2019.

[10] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[11] Arun Das and Paul Rad. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*, 2020.

[12] Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, pages 1–55, 2022.

[13] Andreas Holzinger, Anna Saranti, Christoph Molnar, Przemyslaw Biecek, and Wojciech Samek. Explainable ai methods-a brief overview. In *International workshop on extending explainable AI beyond deep models and classifiers*, pages 13–38. Springer, 2022.

[14] Parvathaneni Naga Srinivasu, N Sandhya, Rutvij H Jhaveri, and Roshani Raut. From blackbox to explainable ai in healthcare: existing tools and case studies. *Mobile Information Systems*, 2022:1–20, 2022.

[15] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9):1–33, 2023.

[16] Adam White and Artur d'Avila Garcez. Measurable counterfactual local explanations for any classifier. *arXiv preprint arXiv:1908.03020*, 2019.

[17] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615*, 2019.

[18] R Caruana. Intelligible models for healthcare. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD'152015.*, page 1721.

[19] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[20] Rasheed Omobolaji Alabi, Mohammed Elmusrati, Ilmo Leivo, Alhadi Almangush, and Antti A Mäkitie. Machine learning explainability in nasopharyngeal cancer survival using lime and shap. *Scientific Reports*, 13(1):8984, 2023.

[21] Aditya Bhattacharya. *Applied Machine Learning Explainability Techniques: Make ML models explainable and trustworthy for practical applications using LIME, SHAP, and more*. Packt Publishing Ltd, 2022.

[22] Gichan Lee and Scott Uk-Jin Lee. An empirical comparison of model-agnostic techniques for defect prediction models. In *2023 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pages 179–189. IEEE, 2023.

[23] Shubham Rathi. Generating counterfactual and contrastive explanations using shap. *arXiv preprint arXiv:1906.09293*, 2019.

[24] Hema Sekhar Reddy Rajula, Giuseppe Verlato, Mirko Manchia, Nadia Antonucci, and Vassilios Fanos. Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. *Medicina*, 56(9):455, 2020.