EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR INTERPRETING MACHINE LEARNING MODELS

Thesis Submitted

in Partial Fulfillment of the Requirements for the Degree of

MASTER OF TECHNOLOGY

IN

DATA SCIENCE

by

Raviraj

(24/DSC/28)

Under the supervision of

Dr. Abhilasha Sharma

Associate Professor, Department of Software Engineering

Delhi Technological University, Delhi, India

To the

Department of Software Engineering

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

MAY, 2026

DEPARTMENT OF SOFTWARE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

I am thankful to the divine for being a pillar of strength, resilience and guidance in this journey. This belief provided clarity and encouragement in the challenges and successes faced throughout the process of this research.

My sincere and heartfelt thanks to my supervisor, Dr. Abhilasha Sharma for her invaluable guidance, constant support and encouragement from the initial stages of this research. Her wise advice, constructive criticism and gentle guidance were instrumental in the direction and quality of this work. Her focus on clarity, rigor and critical thinking has significantly deepened my knowledge and understanding and reinforced my approach to research problems. I am so grateful to her unflagging faith in my capabilities, otherwise this thesis would not have been possible.

I would like to thank Dr. Roshni Singh for her technical inputs, her collaborative approach towards discussions and her constant encouragement throughout the course of my research and M.Tech journey. Her practical insights and thoughtful inputs made a real difference to the development and refinement of key aspects of this work. I am grateful to her for sharing her knowledge and being available to offer guidance when it was needed.

Raviraj
24/DSC/28

DEPARTMENT OF SOFTWARE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, Raviraj, hereby declare that the thesis entitled "Explainable Artificial Intelligence for Interpreting Machine Learning Models", submitted to the Department of Software Engineering, Delhi Technological University, Delhi, in partial fulfilment of the requirements for the award of the degree of Master of Technology, is a record of original work carried out by me under the guidance of the concerned supervisor(s).

I further declare that this thesis has not been copied from any source without proper acknowledgment and citation. This work has not been submitted, either in part or in full, for the award of any Degree, Diploma, Associateship, Fellowship, or any other similar title or recognition at any other institution or university.

Candidate's Signature

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor(s) Signature of External Examiner

DEPARTMENT OF SOFTWARE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE BY THE SUPERVISOR(s)

I hereby certify that the thesis entitled "Explainable Artificial Intelligence for Interpreting Machine Learning Models", submitted by Raviraj (Roll No. 24/DSC/28) to the Department of Software Engineering, Delhi Technological University, Delhi, in partial fulfilment of the requirements for the award of the degree of Master of Technology, is a record of original work carried out by the student under my supervision.

To the best of my knowledge, this work has not been submitted, either in part or in full, for the award of any Degree, Diploma, or similar recognition to this University or any other University/Institution. .

Dr. Abhilasha Sharma
Associate Professor

Place: Delhi

Date:

Department of Software Engineering

EXPLAINABLE ARTIFICIAL INTELLIGENCE
FOR INTERPRETING MACHINE LEARNING MODELS

Raviraj

Abstract

Nowadays, machine learning-based models are becoming increasingly popular in the healthcare environment for disease prediction and creating decision support systems. Although such algorithms usually demonstrate high predictive performance, they tend to be quite black box and, therefore, non-interpretable or at least partially uninterpretable. In case when the system is not interpretable enough, clinicians may not trust it, which ultimately leads to low adoption rates in healthcare environments.

This dissertation introduces an approach to Explainable Artificial Intelligence and creates an explainable machine learning-based model in the context of healthcare. A Random Forest classifier is developed using Breast Cancer Wisconsin Diagnostic dataset to predict binary outcomes – benign vs malignant tumors. To increase transparency and thus improve interpretability, two model-agnostic methods are employed to perform explainability – SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). Both methods are incorporated into the predictive model architecture to provide interpretability while retaining the complexity of the model.

The proposed model achieved a relatively high level of predictive performance, having reached 96.4% accuracy and 96.3% F1-score. SHAP analysis helps to identify global features affecting the prediction and LIME provides local explanations for individual predictions.In the experiments, the outcomes indicate that combining ensemble learning with explainability methods increases predictive accuracy and improves interpretability of the model, at least based on the interpretability perspective.

Contents

Acknowledgement **i**

Candidate's Declaration **ii**

Certificate **iii**

Abstract **iv**

Contents **vii**

List of Table(s) **viii**

List of Figure(s) **ix**

1 Introduction **1**
1.1 Background 1
1.2 Problem Statement 2
1.3 Motivation 2
1.4 Research Objectives 2
1.5 Scope of the Work 3
1.6 Research Contributions 3
1.7 Organization of the Dissertation 3

2 Literature Review **5**
2.1 Introduction 5
2.2 Machine Learning in Healthcare 5
2.3 Explainable Artificial Intelligence 6
2.4 Need for Explainability in Healthcare 6
2.5 SHapley Additive exPlanations (SHAP) 7
2.6 Local Interpretable Model-agnostic Explanations (LIME) 7
2.7 Challenges in Explainable Artificial Intelligence 8
2.8 Random Forest in Healthcare Prediction 9
2.9 Ethical Issues in Explainable Healthcare AI 11
2.10 Research Gap 12

3 Methodology **13**

3.1 Introduction 13
3.2 Framework Overview 13
3.3 Dataset Description 13
3.4 Data Preprocessing 14
3.5 Train-Test Split 15
3.6 Random Forest Classifier 16
3.6.1 Working Principle of Random Forest Classifier 16
3.7 Performance Evaluation Metrics 17
3.7.1 Accuracy 17
3.7.2 Precision 17
3.7.3 Recall 18
3.7.4 F1-Score 18
3.8 Explainability Techniques 18
3.8.1 SHapley Additive exPlanations (SHAP) 18
3.8.2 Mathematical Interpretation of SHAP 19
3.8.3 Local Interpretable Model-agnostic Explanations (LIME) 20
3.8.4 Working Mechanism of LIME 20
3.8.5 Advantages of Explainable AI in Healthcare 21
3.9 System Workflow 22

4 Implementation and Results 23
4.1 Introduction 23
4.2 Implementation Environment 23
4.3 Experimental Setup 23
4.4 Model Training 24
4.4.1 Hyperparameter Configuration and Analysis 24
4.5 Performance Evaluation 25
4.6 Confusion Matrix Analysis 26
4.6.1 ROC Curve Analysis 26
4.7 SHAP-Based Explainability Analysis 28
4.8 LIME-Based Explainability Analysis 31
4.9 Expanded Experimental Results and Interpretation 32
4.10 Clinical Significance of the Proposed Framework 33
4.11 Limitations of the Proposed Framework 34
4.12 Comparative Analysis 34
4.12.1 Additional Comparative Analysis 35
4.13 Discussion 37

5 Conclusion and Future Work 38
5.1 Introduction 38
5.2 Conclusion 38
5.3 Research Contributions 39

5.4 Limitations of the Study 39
5.5 Future Scope 39
5.6 Final Remarks 40
5.7 Social Impact of Explainable Artificial Intelligence in Healthcare 40
5.7.1 Impact on Healthcare Professionals 41
5.7.2 Impact on Patients 41
5.7.3 Ethical and Responsible AI 41
5.7.4 Impact on Society 41

References **43**

List of Table(s)

2.1 Comparative Analysis of Existing Studies 10

3.1 Dataset Description 14

4.1 Implementation Environment 23
4.2 Hyperparameter Configuration for the Random Forest Classifier 25
4.3 Performance Metrics for Random Forest Classifier 26
4.4 Comparative Analysis of Machine Learning Models 35
4.5 Extended comparative analysis of explainable AI models 36

List of Figure(s)

3.1 Proposed Explainable Machine Learning Framework 14
3.2 Workflow of the Proposed Explainable AI Framework 15

4.1 Confusion Matrix for Random Forest Classifier 27
4.2 ROC Curve for Random Forest Classifier 27
4.3 SHAP Summary Plot for Global Feature Importance 29
4.4 LIME Explanation for an Individual Prediction 31

List of Symbols, Abbreviations and Nomenclature

Abbreviation	Meaning
AI	Artificial Intelligence
ML	Machine Learning
XAI	Explainable Artificial Intelligence
SHAP	SHapley Additive exPlanations
LIME	Local Interpretable Model-Agnostic Explanations
RF	Random Forest
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
PDP	Partial Dependence Plot
ROC	Receiver Operating Characteristic
AUC	Area Under Curve
WDBC	Wisconsin Diagnostic Breast Cancer Dataset

CHAPTER - 1

INTRODUCTION

1.1 Background

The advent of Artificial Intelligence (AI), along with other Machine Learning (ML) technologies, has had a remarkable influence on modern healthcare organizations, allowing such innovations as automated diagnosis, predictive analytics, and intelligent clinical decision support systems [1, 2]. Machine learning algorithms can extract hidden patterns from large healthcare datasets and produce highly accurate predictions based on them. Due to the abovementioned advantages, machine learning models have become increasingly popular in healthcare applications, including diagnostics, medical imaging, patient risk prediction, and treatment planning.

In recent years, machine learning algorithms have demonstrated enhanced effectiveness in processing both structured and unstructured medical data, while their performance has remained relatively high [3]. Besides, the use of ensemble learning techniques and deep learning architectures has proved to be very efficient for solving complex problems in healthcare. However, despite high predictive accuracy, many machine learning models are still black-box algorithms, whose inner processes cannot be easily interpreted.

Transparency and interpretability have become paramount issues within healthcare settings as clinical decisions can have significant impacts on patient safety and the results of treatment processes. It is crucial for medical practitioners to be able to receive explanations about the reasons why some prediction was generated by a particular model. Lack of interpretability rapidly undermines trust and, consequently, restricts the application of machine learning algorithms in practice [4].

To solve the above-mentioned issues, explainable artificial intelligence (XAI) has recently become an active field of study [5]. The main goal of XAI is to develop methods which could increase clarity, make reasoning processes more interpretable, and improve the trustworthiness of machine learning models. Explaining machine learning models allows to understand the importance of features, reasoning behind predictions, and behavior of a certain model. Generally, explainability methods provide global explanations which refer to the overall functioning of a model, or local explanations which concern only a single prediction [6].

Among various XAI techniques , SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) are widely adopted because they are model-agnostic in nature and they tend to work well when interpreting complicated machine learn-

ing models [7, 8]. SHAP gives a more broad, like global, feature importance look using ideas from cooperative game theory, and LIME instead tries to clarify each single decision via local surrogate models. This dissertation lays out an explainable machine learning framework for healthcare uses, by merging a Random Forest classifier with SHAP and LIME explanation approaches.

1.2 Problem Statement

The machine learning models have demonstrated quite impressive predictive performance within the field of healthcare. Nevertheless, a large number of highly sophisticated models, such as ensemble learning techniques and deep neural networks, operate in the "black box" mode [9]. They can generate predictions; however, they do not provide sufficient insights into the rationale behind a specific decision.
In the context of healthcare applications, such lack of transparency poses a significant barrier, since clinicians tend to rely on comprehensible input information before accepting any automated predictions. In case of erroneous or unclear results, clinicians' trust can be compromised, which will negatively affect accountability and even cause critical errors. Although a variety of explainability approaches have been proposed, many of the current solutions still fail to balance their ability to provide predictive insights with that of providing interpretable explanations. Consequently, there is a strong need for an explainable machine learning framework that ensures both high predictive performance and meaningful interpretations within the healthcare decision-making process.

1.3 Motivation

The rise in usage of AI-based health care systems has indeed brought about a very high demand for interpretable and reliable machine learning models. Despite the fact that most of the predictive models perform extremely well, their inability to justify the predictions made or how the predictions were made limits their usage in health care settings [10]. Health care practitioners do not easily embrace automated systems if they lack understanding on how the predictions have been arrived at. In case of disease diagnosis, the interpretability of the system becomes critical due to the influence of the decision in determining patients' treatment plans. The main motivation of this project is in some way similar to integrating highly performing machine learning models with explainability approaches. This is aimed at improving reliability and interpretability of health care prediction systems.

1.4 Research Objectives

The main objectives of this dissertation are as follows:

- To analyze interpretability challenges associated with machine learning models in healthcare applications.
- To evaluate existing Explainable Artificial Intelligence (XAI) techniques for improving transparency and trustworthiness in healthcare systems.
- To develop an explainable healthcare prediction model using the Random Forest classifier.
- To integrate SHAP and LIME explainability techniques into the proposed framework for enhancing interpretability and transparency.
- To evaluate the performance of the proposed model using classification metrics such as accuracy, precision, recall, and F1-score.

1.5 Scope of the Work

In any case, the current study focuses on explainable artificial intelligence in supervised machine learning models used in healthcare applications, more or less. In particular, the analysis employs the Breast Cancer Wisconsin Diagnostic (WDBC) dataset to classify the tumor type as malignant or benign using binary classification techniques.
Besides, the scope of the project remains limited to model-agnostic approaches to explain machine learning algorithms, such as SHAP and LIME. Moreover, the focus lies on healthcare datasets with structured numerical data only, without involving deep learning architecture models or real-time healthcare applications.

1.6 Research Contributions

The main contributions of this dissertation can be stated as follows: In the first place, a framework for explainable machine learning in healthcare prediction systems is designed, using the Random Forest classifier as the basis for the prediction system. This framework makes use of SHAP and LIME explainability methodologies in order to obtain both global and local explanations in relation to healthcare predictions. Secondly, a comparative study on Explainable Artificial Intelligence (XAI) techniques is performed in order to assess their performance in healthcare environments. The proposed framework is evaluated through different classification measures with the aim of increasing the transparency, reliability, and robustness of AI-based healthcare systems.

1.7 Organization of the Dissertation

This dissertation is organized into five chapters.

Chapter 1 defines the background, the statement of the problem, the motivation behind the study, the aims and scope of the research, and the contributions made by the study.
Then Chapter 2 goes through a more detailed literature review, covering Explainable Artificial Intelligence techniques in general, healthcare applications, and the interpretability approaches that already exist. The methodology adopted for this project is discussed in Chapter 3. This section not only discusses the data and the preprocessing techniques but also discusses the Random Forest model and the explainability techniques such as SHAP and LIME.
Experimental set up including the evaluation measures, performance analysis, explainability analysis, and result discussions can be found in Chapter 4.
Chapter 5 concludes the entire report by discussing the findings, limitations, and future directions of research.

CHAPTER - 2

LITERATURE REVIEW

2.1 Introduction

Explanation of Artificial Intelligence (XAI) has become an interesting topic for research due to the increased use of machine learning algorithms in sectors that require confidentiality, such as healthcare, finance, cybersecurity, and autonomous systems [5]. Although numerous machine learning algorithms exhibit impressive levels of predictive performance, most of the recent approaches remain as black-boxes that make it difficult to understand the decision-making process of the underlying algorithm [11]. When transparency is missing, then trust, accountability, and even fairness end up being more complicated, and it becomes harder to rely on the results, especially in real world settings.

To address these issues, researchers suggested a range of interpretability techniques, aiming to increase transparency and to characterize model behavior more clearly. In general, these approaches can be grouped into intrinsic interpretability methods and post-hoc explainability methods [6]. Intrinsic models are interpretable by their very design, while post-hoc techniques are meant to rationalize models after they are already trained, not before.

In this chapter, a thorough review is provided of the literature on explainable artificial intelligence, machine learning interpretability, healthcare use cases, SHAP, LIME, and other related explainability strategies. Tonekaboni et al. also stressed that clinicians need context sensitive and interpretable explanations before they consider using AI- assisted clinical decision support systems [12].

2.2 Machine Learning in Healthcare

Machine learning has, in a big way, improved healthcare systems , by enabling disease prediction, medical image analysis, clinical diagnosis,and patient risk assessment [2]. In practice, ML models can sift through large-scale medical datasets and find non-obvious patterns that, may not be easily observable by clinicians.

A few studies show how well machine learning works in healthcare use cases . Jordan and Mitchell [1] talked about the rising significance of machine learning across multiple areas, including healthcare and medical analytics. Ensemble learning approaches, and deep learning architectures too, have demonstrated solid predictive ability for disease diagnosis and treatment recommendation systems .

However, even with their high predictive accuracy, quite a lot of healthcare ML systems still

have weak interpretability. Clinicians need explanations that actually make sense, before they will trust automated predictions [4]. So explainability has turned into a key requirement for healthcare AI systems , in general.
Also, newer deep learning and graph convolutional approaches have pushed intelligent activity recognition further along,and they've supported healthcare-related AI applications as well [13].

2.3 Explainable Artificial Intelligence

Explainable Artificial Intelligence (XAI) is about techniques, and methods that make machine learning models more transparent, and easier to interpret [5]. With XAI, users can get a sense of how the model actually forms predictions , and also which factors seem to steer the decision-making, or at least contribute to it. According to Guidotti et al. [6], explainability approaches can be grouped into a few categories like this:

- Intrinsic Interpretability Methods
- Post-hoc Explainability Methods
- Global Explanation Methods
- Local Explanation Methods

Intrinsic interpretability is typically tied to models that are, in a sense, interpretable by nature e.g. decision trees, linear regression, and rule-based systems. The post-hoc family is used after the model has been trained, so it explains predictions that come from heavier black-box models.Recent studies have emphasized that explainability methods are essential for improving transparency, accountability, and fairness in machine learning systems deployed in critical domains [14]. Global explanation methods attempt to represent the overall behavior of the model. Local explanation methods focus on explaining why a certain prediction is made. [15]. Samek et al. also highlighted that understanding and visualizing deep learning models matters a lot, for better transparency and interpretability across AI systems [16]. Recent developments in interpretable machine learning have made AI models more transparent and accountable [17, 18].

2.4 Need for Explainability in Healthcare

The field of healthcare is considered a high-stake domain wherein any erroneous prediction can lead to significant downstream implications. Thus, transparency and interpretability become almost mandatory necessities when using AI to assist clinical systems [9]. Clinicians need to be informed clearly about the reasoning behind the model's diagnostic decision. Explainability helps clinicians to verify the results, identify important features, and increase confidence in AI-powered systems.

Holzinger et al. [4] argued that an explainable AI system is an essential requirement in the medical field since clinicians should comprehend the reasoning of the algorithm before utilizing its predictions. Similarly, Goodman and Flaxman [10] discussed the importance of transparency, as well as accountability, in AI-powered decision-making systems.

2.5 SHapley Additive exPlanations (SHAP)

SHapley Additive exPlanations (SHAP) is among the most commonly used techniques for interpreting machine learning models [7]. SHAP somewhat relies on game theory and defines each feature's contribution based on the influence on the prediction. The SHAP method exhibits a number of theoretical properties including consistency and local accuracy. Therefore it will not only provide you global but also local interpretability as values are computed on a prediction-by-prediction basis, but then averaged for a whole dataset in a more readable way than one single perspective. Interpretation techniques using feature attribution have been proven to be efficient to discover driving variables for predictions on complicated data from healthcare [19].
The SHAP values have been used to implement an integral interpretation framework designed by Lundberg and Lee [7]. The SHAP summary plots are frequently employed to demonstrate global feature importance and to show how features affect predictions directly. However, even with all these advantages, the approach is computationally expensive, particularly when dealing with massive amounts of data and a sophisticated model. Nevertheless, thanks to the sound theory behind it and high interpretability provided by SHAP, the method is rather successful in healthcare settings.

2.6 Local Interpretable Model-agnostic Explanations (LIME)

Local Interpretable Model-agnostic Explanations (LIME) is a well known post-hoc explainability method proposed by Ribeiro et al. [8]. LIME tries to explain single predictions by approximating the behavior of a complicated model, using an interpretable local surrogate model. The way it works is by slightly perturbing input samples, then watching what happens to the prediction scores, so it can infer which features contribute more. In practice it is model-agnostic , so it works with pretty much any machine learning approach, without needing to touch or modify the original model, or at least that's the idea.
LIME gives local explanations that are handy for grasping specific outcomes, especially in healthcare systems. Clinicians may inspect how certain medical features steer diagnosis predictions, for a particular patient, not for the whole population. Still, even if the method increases interpretability, the resulting rationales can become a bit unstable when perturbations are small or when the dataset changes.

2.7 Challenges in Explainable Artificial Intelligence

Despite the fact that Explainable Artificial Intelligence (XAI) has already managed to significantly enhance the transparency and interpretability of machine learning algorithms, there are some obstacles in terms of creating reliable and trustworthy explainable models. This is relevant in the healthcare context, where choices made in the treatment of patients affect patient wellbeing.The problems of explanation stability, fairness and robustness have been pointed out in multiple studies with respect to healthcare AI systems [20, 21].
The computational complexity is one of the major hurdles in the field of Explainable AI. For some explainable methods, like SHAP, multiple computations of feature contributions are required. It may be expensive to compute, so may not be feasible in some application use-cases (e.g. Live healthcare decisions). These challenges with explainability techniques are also noted in other research fields such as computational complexity or the consistency of explanations and scalability to large datasets in healthcare [22].
However, as it was mentioned above, interpretability is subjective, and it can change dramatically based on the level of knowledge possessed by a particular individual. Thus, the interpretation which is understandable to data scientists may not be comprehensible for the clinician. Consequently, creation of interpretations which will be user-friendly and at the same time clinically significant remains a tough nut to crack. Another challenge which arises repeatedly is the matter of global versus local interpretability balance. While global interpretability methods aim to provide a comprehensive understanding of how the model works, local interpretability approaches focus on a specific prediction only. Nevertheless, providing local explanations in addition to a comprehensive one is quite a difficult task. Bias and fairness also remain the key concerns for Explainable Artificial Intelligence systems. For example, if the training set contains a bias, the generated explanations will reflect this characteristic as well. This might lead to incorrect predictions which will influence the quality of patient care. The last but not least important issue that should be addressed by healthcare experts is the question of privacy and security. The other difficulty in this area comes from attempts to assess the quality of explanations. At the moment, there is no universal criteria for assessing the understanding, reliability, or clinical relevance of explanations. Different approaches to explaining may offer conflicting assessments of the same prediction, making evaluation somewhat complicated. For example, criteria for "good" explanations may vary slightly depending on the point of view one takes, making direct comparison difficult.
However, Explainable Artificial Intelligence remains a valuable tool in improving transparency, accountability, and trustworthiness of machine learning models. Nevertheless, addressing these limitations becomes crucial if we want to develop clinically acceptable artificial intelligence-assisted healthcare applications.

2.8 Random Forest in Healthcare Prediction

Random Forest, in essence, is an ensemble learning technique that was developed by Breiman [23]. It combines several decision trees, thereby increasing predictive accuracy and decreasing overfitting simultaneously. Random Forest models are highly effective in structured tabular healthcare datasets since they can capture both non-linear relationships between features and different types of interaction among them. In addition, there is this additional output, namely feature importance metrics, which essentially serve to identify important variables. Several works have applied Random Forest classifiers in healthcare scenarios, such as in the context of disease classification and patient risk prediction.

Table 2.1: Comparative Analysis of Existing Studies

Study	**Method Used**	**Application Area**	**Limitation**
Jordan and Mitchell [24]	Machine Learning	Healthcare Prediction	Limited interpretability
Ribeiro et al. [25]	LIME	Local Explainability	Instability in explanations
Lundberg and Lee [26]	SHAP	Global Explainability	High computational complexity
Guidotti et al. [27]	XAI Survey	Black-box Models	Limited practical evaluation
Doshi-Velez and Kim [28]	Interpretable ML	Explainability Evaluation	Lack of standard metrics
Adadi and Berrada [29]	Explainable AI Review	Transparency Analysis	Scalability challenges
Lipton [30]	Interpretability Analysis	AI Transparency	Conceptual ambiguity
Samek et al. [31]	Deep Learning Explainability	Neural Networks	High computational overhead
Tjoa and Guan [32]	Medical XAI Survey	Healthcare Systems	Limited deployment studies
Caruana et al. [33]	Intelligible ML Models	Pneumonia Prediction	Scalability limitations
Murdoch et al. [34]	Interpretable ML	General AI Systems	Limited healthcare validation
Bhatt et al. [35]	Explainable ML Deployment	Real-world AI Systems	Robustness concerns
Vilone and Longo [36]	XAI Evaluation Methods	Interpretability Assessment	Lack of unified framework
Nauta et al. [37]	Healthcare XAI	Medical Diagnosis	Limited benchmark datasets
Zhang and Chen [38]	XAI Review	Healthcare AI	Computational cost
Ali and Khan [39]	Healthcare Explainability	Clinical Applications	Ethical challenges
Bussone et al. [40]	Clinical Decision Support	Healthcare Informatics	User dependency issues
Goodman and Flaxman [41]	Transparent AI Systems	Regulatory AI Frameworks	Limited technical implementation
Linardatos et al. [42]	Interpretability Review	Machine Learning Models	Limited experimental comparison

2.9 Ethical Issues in Explainable Healthcare AI

However, the increasing application of artificial intelligence within healthcare has raised numerous ethical issues, including those of transparency, fairness, accountability, privacy, and general trustworthiness. Even in instances where explainable artificial intelligence can be applied successfully, for example, enhancing the interpretability of machine learning algorithms, the ethical implications in AI-assisted healthcare cannot be ignored. It is, therefore, difficult to distinguish between these issues.

Firstly, the issue of ethical violation in artificial intelligence revolves around patient privacy and confidentiality. Most of the time, the data used in healthcare is sensitive and involves the personal medical record of patients, including diagnosis, treatments administered, and other health-related features. Such data should be managed properly, lest it leads to privacy and data confidentiality breaches. Fairness and bias are two ethical issues that need to be considered while creating an algorithm in AI-based healthcare. The inclusion of unequal data in the training dataset may cause biases in the model, which, in turn, may lead to biased predictions about certain patient groups. This will affect the accuracy of diagnosis, influence the treatment process, and even impact the availability of healthcare services for certain patients.

Accountability is another essential ethical problem related to AI-generated decisions. In particular, mistakes made by AI can lead to poor treatment options for patients in the healthcare setting. This means that, in order to ensure proper care and treatment, it becomes crucial to determine responsibility and accountability of the person making the choice based on AI algorithms.Transparency and interpretability of AI-generated decisions also become important from the ethical point of view. It is crucial that clinicians understand the logic behind AI-generated predictions in order to use automated decision-making technologies.

Explainable AI techniques such as SHAP and LIME assist in the push towards transparency by providing more understandable explanations for model outputs. Ethical considerations related to informed consent apply to the use of healthcare artificial intelligence. Patients should always be notified when artificial intelligence is applied in diagnosis or in decision-making regarding treatment options. Maintaining transparency regarding the use of artificial intelligence promotes trust and acceptance of the process, even if it feels somewhat "modern."

Security risks and cyber threats in general raise ethical concerns in regard to healthcare artificial intelligence systems. The situation can become complex since any unauthorized access results in adversarial attacks as well as data manipulation, which can decrease prediction accuracy and negatively impact patient safety. Therefore, secure implementation and verification are crucial for the development of healthcare AI.

However, regulatory compliance and ethical governance are key considerations when implementing the use of explainable health care systems. Health care AI systems should comply with relevant regulations, ethics, and other aspects related to the field so that the implementation process remains responsible and safe. Furthermore, while ethical governance may take

some time, it must occur.
Nevertheless, despite the above ethical issues, Explainable Artificial Intelligence technology also has a good potential for ensuring that health care systems become more transparent, responsible, accountable, equitable, and credible in general. Dealing with ethical issues at the very beginning is important if one wishes to develop AI-enabled health care systems responsibly.

2.10 Research Gap

Even with the rapid advancement of Explainable Artificial Intelligence (XAI), there are still a few persistent problems when it comes to building transparent and reliable healthcare prediction systems. A lot of the usual machine learning models in healthcare end up behaving like black-box systems, and then it becomes hard, for clinicians to grasp why a certain prediction happens in the first place.
Most of the existing work has tended to lean either toward strong predictive accuracy, or else toward interpretability improvements in a sort of separate lane. Still, keeping a workable equilibrium between predictive performance and explainability is, in practice, one of the major hurdles for healthcare AI systems.
Moreover, many explainability methods tend to provide either local explanations, or global explanations, but separately. For example, LIME is known for providing local explanations for one specific prediction, whereas SHAP focuses on providing global feature importance analysis. However, there is a lack of research related to the integration of both local and global explainability in one consistent healthcare prediction framework.
Rudin stated that more interpretable models must be preferred over black-box models, in case of critical applications such as in the field of healthcare, as mentioned in [43].
Additionally, most current techniques have shortcomings such as unstable explanations, high computational complexity, limited clinical interpretability, and poor scalability when dealing with large healthcare data sets. Even some approaches fail to provide necessary visualizations and comparison to understand the working of the model itself, which, in fact, is the core purpose of such techniques, yet it remains underdeveloped. Furthermore, many techniques do not consider ethical issues such as fairness, accountability, bias detection, and trustworthiness in their prediction models. Consequently, it appears that there is a real need for an explainable healthcare prediction framework that combines strong predictive ability with both global and local interpretability techniques.

CHAPTER - 3

METHODOLOGY

3.1 Introduction

The chapter discusses the proposed explainable machine learning framework for healthcare prediction systems. This framework integrates a Random Forest classifier with XAI techniques to enhance the predictive performance of the algorithm and increase its transparency. Overall, the methodology is associated with the diagnosis of breast cancer based on the Breast Cancer Wisconsin Diagnostic (WDBC) dataset.
Specifically, the methodology comprises such steps as data preprocessing, modeling, generating the prediction outputs, and performing an explainability analysis (SHAP + LIME). The purpose of this approach is to offer practical insights into the operation of the model and make its reasoning transparent and easy to apply in practice.

3.2 Framework Overview

As per the proposed approach, the process can be described as being somewhat composed of a number of phases such as data collection, preprocessing, model learning, prediction, and finally explainability, among others. The overall process is illustrated in Fig. 3.1.
In the beginning , the framework preprocesses the dataset and trains a Random Forest classifier by using structured healthcare data. Then after prediction generation is done, explainability techniques including SHAP and LIME are used to give both global explanations and local ones, so you can interpret the behavior better.

3.3 Dataset Description

The Breast Cancer Wisconsin Diagnostic (WDBC) dataset shows up in this dissertation for binary classification, of breast cancer tumors. This dataset is publicly available and widely used for many healthcare prediction studies, so people keep reusing it. The data includes 569 instances, and also 30 numerical features that come from digitized images of fine needle aspirate (FNA) tests. These features capture characteristics of cell nuclei and morphology.
The dataset also contains several numerical healthcare features extracted from digitized images of fine needle aspirate (FNA) tests. Among the most important ones are radius, texture, perimeter, area, smoothness, compactness, concavity, symmetry, and fractal dimension, and those are used as indicators for separating malignant from benign tumors.

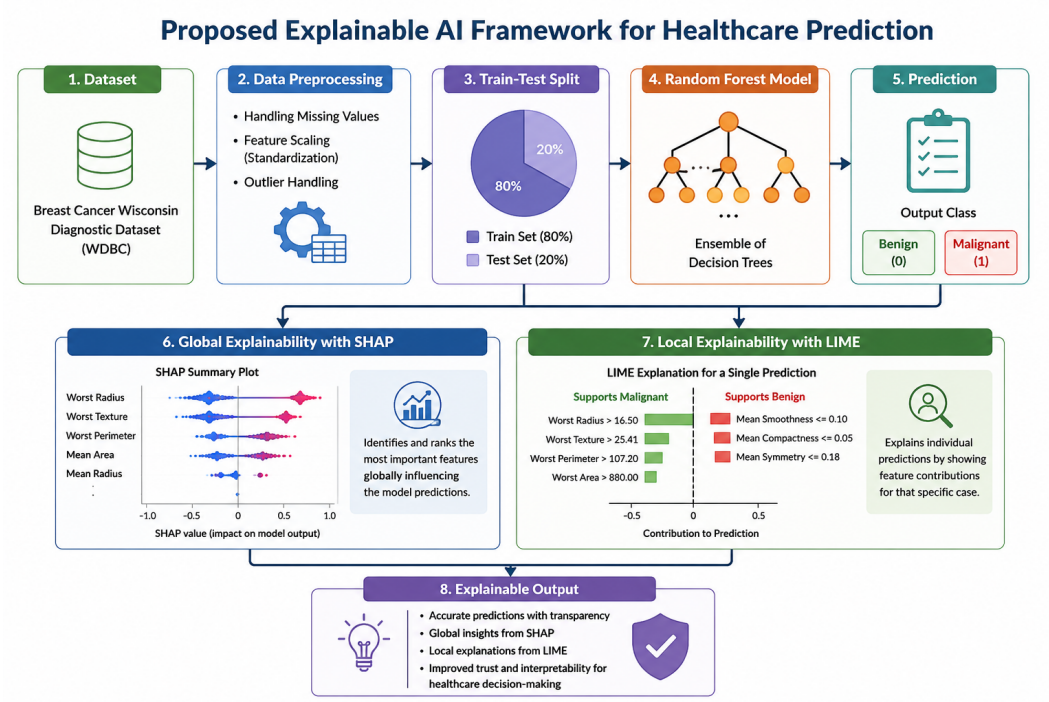

Figure 3.1: Proposed Explainable Machine Learning Framework

The target variable has two classes which are labeled as:

- Malignant Tumor (1)
- Benign Tumor (0)

Table 3.1 provides the dataset overview, and the main counts.

Table 3.1: Dataset Description

Parameter	**Value**
Dataset Name	Breast Cancer Wisconsin Diagnostic
Number of Samples	569
Number of Features	30
Classes	Benign, Malignant
Feature Type	Numerical
Missing Values	None

3.4 Data Preprocessing

In this work, a preprocessing pass is done so the dataset ends up cleaner and also is ready for the machine learning stage. During this part, we cover data cleansing, label encoding,

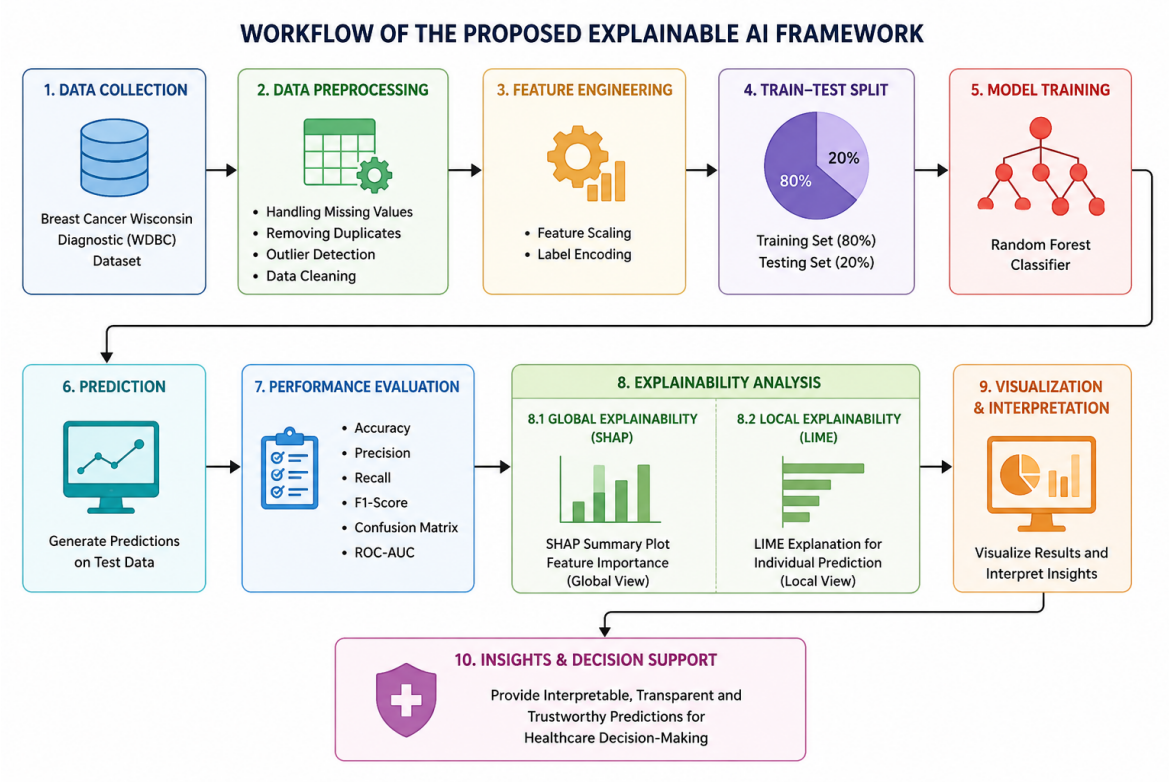

Figure 3.2: Workflow of the Proposed Explainable AI Framework

feature normalization, and the usual train-test split. At the beginning, the unnecessary identifier columns get dropped, because they don't help in the learning process. Next comes the conversion of target labels into numeric form in the context of binary classification. Subsequently, the process of feature scaling through standardization takes place to make sure that the magnitude of all features becomes somewhat comparable and the model exhibits stability. In actual practice, standardization involves rewriting the value according to the formula given below:

$$Z = \frac{X - \mu}{\sigma} \tag{3.1}$$

where,

- X denotes the raw value of the attribute
- μ denotes the mean value estimate
- σ denotes the standard deviation parameter

3.5 Train-Test Split

The data set is then divided into training and testing data sets in the ratio 80:20. Stratified sampling is employed to maintain the distribution of classes approximately constant in both

data sets despite the extra effort involved. The training data set is used to build the model while the testing data set is employed for validation purposes.

3.6 Random Forest Classifier

Random forest is a type of ensemble learning algorithm introduced by Breiman [23]. In reality, it constructs a collection of decision trees and then blends their outcomes, meaning that accuracy increases, while overfitting decreases. The random forest approach appears to perform well with structured data in healthcare due to its ability to detect nonlinear relationships between features patterns, not only straight associations.Apart from that, the model automatically provides feature importance scores that can be used to detect important variables, not too hard.
In this project, the Random Forest classifier will be configured as shown below: The classifier will use 100 decision trees, and Gini impurity will be used as the criterion for splitting the nodes.
The Gini impurity employed for node splitting is computed as

$$Gini = 1 - \sum_{i=1}^{n} p_i^2 \tag{3.2}$$

where p_i denotes the probability of a class at a node.

3.6.1 Working Principle of Random Forest Classifier

Random Forest is an ensemble based machine learning method , where it brings together several decision trees to make predictions better and also, to lower the overfitting risk. It works on the idea of ensemble learning, so in practice a bunch of weaker learners get mixed together, and somehow you end up with a more reliable predictive model.
In the Random Forest classifier , many decision trees get built during training. Then, when the time comes to predict, the model looks at what each tree says and uses majority voting , to pick the most common class. Each tree is not trained on exactly the same information either. Instead it uses a randomly chosen slice of the training set , plus a random selection of features. This randomization also assists in achieving diversity in models, which in turn improves the overall stability of the forest.
The process of Random Forest Classifier can be depicted as follows:

1. Firstly, use bootstrap sampling to generate random samples of the training dataset.
2. Then, train a decision tree separately for each subset.
3. During the training of the decision tree, while splitting nodes, consider only a subset of random features and not all at once.
4. Following this step, a decision tree makes predictions for the input datasets.

5. finally, class label is assigned by majority voting of all trees in the forest.

So it has this ensemble type structure and it controls the variance and increases the generalization accuracy compared to a single decision tree. Because there are lots of trees voting for one prediction the model is more stable and not swayed by the noisy features. The Gini impurity is used by random forest classifier in splitting the useful features. Gini impurity is defined like this:

$$Gini = 1 - \sum_{i=1}^{n} p_i^2 \tag{3.3}$$

where:

- p_i is the probability of class i
- n is the total count of classes

Regarding practical aspects, Random Forest offers several advantages when applied to healthcare predictions. First, it allows handling high-dimensional medical data effectively, mitigates the problem of overfitting, detects feature interactions, and yields good predictions. In addition, feature importance values estimated using Random Forest can improve model interpretability and reveal those medical features that contribute to disease prediction. Due to the aforementioned reasons, Random Forest finds application in healthcare-related areas such as disease diagnostics, cancer prediction, medical risk assessment, and clinical decision-making systems.

3.7 Performance Evaluation Metrics

The predictive capabilities of the model are evaluated using different classification measures such as accuracy, precision, recall, and even the F1 score, which can be quite important at times, to be honest.

3.7.1 Accuracy

The accuracy measures how right the model is in general – in some sense a global measure of correctness of predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.4}$$

3.7.2 Precision

The precision is the proportion of predicted positive cases which are actually positive – or, in other words, a purity of positive predictions.

$$Precision = \frac{TP}{TP + FP} \tag{3.5}$$

3.7.3 Recall

The recall (also called sensitivity, roughly speaking) shows how well the model captures positive cases.

$$Recall = \frac{TP}{TP + FN} \tag{3.6}$$

3.7.4 F1-Score

The F1-score is basically the harmonic mean between precision and recall, some people call it a compromise really, because it balances both. It is computed as this below

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3.7}$$

where i n general , the following symbols are used :

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives.

3.8 Explainability Techniques

Explainability techniques are integrated into the proposed framework to improve model transparency and interpretability.

3.8.1 SHapley Additive exPlanations (SHAP)

SHAP is a model-agnostic explainability technique based on cooperative game theory[7] . kind of, it gives contribution values to features depending on how much they sway the predictions
In practice, SHAP values are obtained from this equation:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)] \tag{3.8}$$

where:

- ϕ_i stands for the SHAP value tied to feature i
- F is the overall feature set
- S denotes particular feature subsets, or partial groupings

- $f(S)$ means the model prediction computed with S

With SHAP, you get both global readouts and local interpretability , because it pinpoints which features are doing the heavy lifting for the model outputs, at least according to the contribution scores.

3.8.2 Mathematical Interpretation of SHAP

SHapley Additive exPlanations (SHAP) is some kind of Explainable Artificial Intelligence method, based on cooperative game theory. SHAP helps explain machine learning outcomes by giving contribution scores to individual input features , depending on how much they affect the model output, not just in a rough way, but in a kind of systematic manner.
The SHAP setup comes from the idea of Shapley values that were first brought in by game theory. Here, every feature is treated like a player that takes part in a cooperative game , while the prediction from the machine learning model is viewed as the reward that the players basically share (or compete for).
The SHAP value for a feature is the average marginal contribution of that feature across all possible arrangements of feature subsets. so, attribution stays fair and also consistent when you compare how important different input variables are for the prediction.
Mathematically, the SHAP value for feature i is computed as:
The equation above figures out the contribution of feature i by checking how the prediction shifts when that feature gets added to various groups of features, each group being a different subset.
The ability of SHAP to provide both local and global interpretations is one of the major advantages of this method since one can analyze the results on both levels.
Thus, global explanations are used to understand which features have an approximate influence on the model behavior, whereas local explanations refer to analyzing individual predictions made for particular patients. When using SHAP in healthcare settings, medical professionals will be able to understand how certain clinical factors, such as tumor radius, texture, area, and concavity, influence disease predictions in practice. With the help of SHAP summary plots and feature importance visualizations, one gets clear signals from the model.
Another significant property is consistency. It is actually an important feature in this case since SHAP values assigned to a particular feature will increase together with its importance. That is, any changes in the impact of some feature on the model will result in a proportional change of SHAP values, which is important since this technique is considered more reliable compared to several other explainability methods.
However, there are still some disadvantages to the SHAP approach. If we consider the case of big data and an extremely complicated machine learning model, the SHAP method might require significant computational power. However, due to its strong theoretical base and ability to provide explanations for predictions, SHAP is widely applied within AI healthcare tools.

3.8.3 Local Interpretable Model-agnostic Explanations (LIME)

LIME is the local method for explaining predictions, proposed by Ribeiro et al. [8]. Essentially, this approach seeks to explain each prediction using the substitution of an opaque model with an interpretable surrogate model, somewhat similarly, yet not precisely the same. LIME involves sample perturbation followed by observation of changes in predictions, from which inference is drawn about influential features.
Optimization formulation used within LIME is expressed below:

$$\xi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g) \tag{3.9}$$

The terms in this formula represent:

- f - the original model
- g - the interpretable model (surrogate model)
- π_x - the locality measure
- $\Omega(g)$ - the complexity measure of the model

With LIME, instance-level explanations are available, which may prove useful for interpretation of specific output within healthcare systems, despite the classifier itself being less interpretable.

3.8.4 Working Mechanism of LIME

Local Interpretable Model-agnostic Explanations (LIME) is a method in the field of Explainable Artificial Intelligence that seeks to interpret individual predictions generated by complex machine learning models. Unlike the methods associated with global explainability, LIME tends to focus more on understanding the behavior of a model in the vicinity of a particular prediction rather than its general characteristics.
The process through which LIME works first involves selecting one particular instance of the dataset, which is used as the basis of this interpretation. This instance is followed by generating a number of perturbed samples in its proximity by changing their feature values slightly. These samples are used as inputs to the already existing machine learning model to obtain prediction outputs as well as the details of what occurred during their generation.
Upon having collected all of the required prediction outputs, LIME assigns weights to each of the perturbed samples based on their similarity to the original instance. Therefore, samples that are closer to the original data point receive higher weights, whereas the ones that are farther away are assigned smaller weights. And after that, a simpler model is created locally. This locally interpretable model is leveraged to approximate the decision boundary that exists near the input data point. From this approximation, the locally interpretable model reveals those features that are important, and thus, the importance of individual features in the decision-making process.

LIME works by minimizing the following objective function, and I think it pretty much sums up the goal of LIME, at least for now.
The objective of LIME is to provide explanations that are both locally accurate and easily understandable to human beings. When it comes to healthcare prediction systems, LIME can be employed to explain individual diagnosis results. Clinicians can be explained how their patients are diagnosed by the algorithm and which are the most important clinical attributes that influenced the diagnosis, high-risk or low-risk.
As an illustration, in a breast cancer prediction system, LIME can identify the most significant factors that affect the prediction, such as tumor radius, texture, concavity, and smoothness.
The primary benefit of LIME is its flexibility, considering that the technique can be applied to various machine learning models without modifications to the existing model. However, LIME explanations may still vary slightly depending on the use of random perturbation sampling. Nevertheless, LIME continues to appear as one of the most popular approaches for explaining models in interpretable machine learning and healthcare applications.

3.8.5 Advantages of Explainable AI in Healthcare

Explainable Artificial Intelligence (XAI) appears to play an important role in promoting the adoption of machine learning systems in the healthcare setting. On the one hand, black box models yield extremely precise predictions. However, on the other hand, their inability to provide insight into the process makes it difficult for clinicians and other health professionals to use them. XAI approaches, in effect, make model behavior more transparent, thus increasing confidence in AI assisted decision support systems.
Transparency is one of the main benefits of Explainable Artificial Intelligence. Clinicians can better understand the reasons behind the emergence of certain diagnosis or prediction, which allows them to verify the consistency of the output against medical knowledge and clinical observations. In other words, instead of receiving only the output, the user gets to know the reasoning trail, which is exactly what they need.
The use of explainability techniques also increases trustworthiness and accountability within healthcare organizations. In critical medical situations, incorrect predictions may be very dangerous for patients' lives and well-being. The provision of easily interpretable explanations reduces the uncomfortable gap of uncertainty and makes the automated system more trustworthy for its users. This will enable healthcare professionals to better understand patterns of disease and make decisions on evidence, rather than an unknown score which no one can comprehend.
Local methods of explainability like LIME can offer something that can be useful in the diagnosis of individual patients. They are able to give some sort of local explanation for the operation of the model, and they can explain the "decision process" regarding some particular patient. These explanations can improve interpretability of models and make communication between clinicians and AI easier.

Finally, Explainable AI can significantly help in compliance with regulations and ethics principles of AI. When there is transparency in a machine learning system, then its validation becomes much easier. Thus, when we combine explainability methods with machine learning models, we will significantly enhance transparency, interpretability, trust, and clinical usability of the healthcare prediction systems.

3.9 System Workflow

And here is the general idea of the workflow of our suggested framework: first comes the loading of the healthcare dataset, then preprocessing, and after that feature scaling, you got it all, just the standard procedure. Once it has been processed the dataset gets split into training and testing subsets , using stratified sampling. Then the Random Forest classifier is trained on the training data and for the test samples predictions are produced. After that, predictive quality is checked with common classification measures such as accuracy, precision, recall , and the F1-score. For better interpretability, SHAP is brought in for global explainability, while LIME is used to craft local explanations for each individual prediction. Finally, all explainability results are visualized and interpreted, so the transparency and trustworthiness in healthcare decision systems can be improved over time.

CHAPTER - 4

Implementation and Results

4.1 Introduction

This chapter presents the implementation details, the experimental setup, performance evaluation, some explainability analysis, and then a discussion of the results that came out from the proposed explainable machine learning framework. Overall, the framework is put together using Python, plus different machine learning libraries. The Random Forest classifier is trained on the Breast Cancer Wisconsin Diagnostic dataset, while SHAP and LIME techniques are used in order to improve interpretability and also transparency, of predictions.

4.2 Implementation Environment

The implementation of the proposed framework was done with the Python programming language, plus open-source machine learning libraries. The Table 4.1 shows what I used , in terms of software and hardware specifics for this work.

Table 4.1: Implementation Environment

Component	**Specification**
Programming Language	Python 3.10
Development Environment	Jupyter Notebook
Machine Learning Library	Scikit-learn
Explainability Libraries	SHAP, LIME
Visualization Libraries	Matplotlib, Seaborn
Operating System	Windows 11
Processor	Intel Core i5/i7
RAM	8 GB or Higher

4.3 Experimental Setup

The breast cancer wisconsin diagnostic data set is partitioned into training and test sets, in an approximate ratio of 80% to 20%. In order to ensure that there is not much difference in the distribution of labels, the method of stratified sampling is employed such that both the

subsets have similar proportions of classes. The random forest model is then fitted with 100 trees using the Gini impurity criterion for evaluating the quality of split.
The overall approach can be depicted as follows

1. Dataset loading
2. Data preprocessing
3. Feature scaling
4. Train test splitting
5. Model training
6. Prediction generation
7. Performance evaluation
8. SHAP analysis
9. LIME analysis

4.4 Model Training

The Random Forest classifier is trained using the preprocessed healthcare data set, and for some reason, the ensemble learning technique enables the classifier to detect complex relationships between features, hence improving its predictive performance. The model is trained using the information present in the training data set and generates predictions for the new testing data instances that have never been seen before by the model. In the end, the trained model performs well when classifying malignant tumors against benign ones.

4.4.1 Hyperparameter Configuration and Analysis

The performance of machine learning models is somewhat highly dependent on the hyperparameter tuning. Hyperparameters are predetermined values that govern the learning process of the model and can affect the predictive performance, generalization ability, and computational efficiency. Hence, incorrect hyperparameter tuning can greatly impair the stability, which makes it necessary to apply proper hyperparameter tuning for creating stable health care predictive models.
In this dissertation, a Random Forest classifier is trained using various hyperparameters that allow achieving the desired classification results on the Breast Cancer Wisconsin Diagnostic dataset. These hyperparameters are experimentally selected based on the need to ensure both high predictive quality and model stability. In the case of the developed Random Forest classifier, the number of decision trees is set to 100. With an increase in the number of trees, the stability of predictions improves and the variance of the model decreases. Nevertheless,

with a further increase in the number of decision trees, computational complexity may grow while performance remains almost unchanged. Thus, 100 decision trees were used to train the model.
Gini impurity criterion is utilized in splitting nodes when constructing the tree. The Gini impurity criterion shows, in a sense, the probability of making an error during the classification at a particular decision node, thus allowing determining which split of the feature is optimal. With it, one will be able to split the healthcare features in an optimized way, and yes, it enhances the classification performance as well.The train test split ratio is set to 80:20 using the stratified sampling method. Stratified sampling preserves the same class distribution for malignant and benign tumors in both datasets: the training and the testing ones. Such an approach usually increases evaluation credibility and prevents the class imbalance issue from appearing.
A specific random seed is chosen to make the experimental results reproducible. Reproducibility is important for developing AI healthcare solutions because stable model performance is necessary in clinical decision support applications. The features are scaled through standardization to normalize their values prior to the model training. The maximum depth for Random Forest classifier model remains unbound, which allows it to capture all the complexities of feature interaction in health care data. In case the maximum depth is set too low, it could reduce the prediction capability of the algorithm. However, when the maximum depth is set too high, the algorithm absorbs non-linearities in a more clear manner.
Table 4.3 illustrates important hyperparameters chosen for the proposed Random Forest classifier model.

Table 4.2: Hyperparameter Configuration for the Random Forest Classifier

Hyperparameter	**Value**
Number of Trees (n_estimators)	100
Criterion	Gini Impurity
Train-Test Split Ratio	80:20
Random State	42
Feature Scaling	Standardization
Maximum Depth	None

Using this design, the suggested approach will be able to achieve accurate predictions while being stable and interpretable at the same time. Moreover, tuning hyperparameters properly is important for the generalization capability of the model and reduces the chances of overfitting, which is crucial for medical prediction systems.

4.5 Performance Evaluation

The predictive accuracy of the framework is evaluated through measures such as accuracy, precision, recall, F1 score, and error rate as well.

The performance metrics for the Random Forest model are presented in Table 4.3.

Table 4.3: Performance Metrics for Random Forest Classifier

Metric	Value	Interpretation
Accuracy	0.964	High overall correctness
Precision	0.972	Low false positives
Recall	0.955	High true positive rate
F1-Score	0.963	Balanced performance
Error Rate	0.036	Low misclassification

From the findings above, the proposed framework appears to possess satisfactory predictive capabilities for virtually all metrics used. A high precision value indicates that the system produces fewer false positive predictions, while maintaining a high recall value means that the malignant cases are detected effectively. In light of this, it becomes evident that the F1-score is accurate, and hence the approach can be effectively utilized within the healthcare domain due to the importance of both false positive and negative cases.

4.6 Confusion Matrix Analysis

This is precisely what the confusion matrix does by presenting a fairly detailed view of classifier performance through a juxtaposition of the predicted and actual labels.
As illustrated in Figure 4.1, majority of samples turn out to be accurately classified, while the number of misclassification errors remains relatively low for both classes. The classifier appears to maintain consistency in its predictions, showing no signs of bias towards any class.It is also worth to note that the number of false negatives is very limited, which can have very serious consequences in the field of health care, since the diagnosis of a malignant case may go undetected.

4.6.1 ROC Curve Analysis

The ROC curve is used to assess the classification performance of the trained Random Forest classifier.The classifier achieved a great AUC score indicating great distinction between both classes which were malignant and benign.

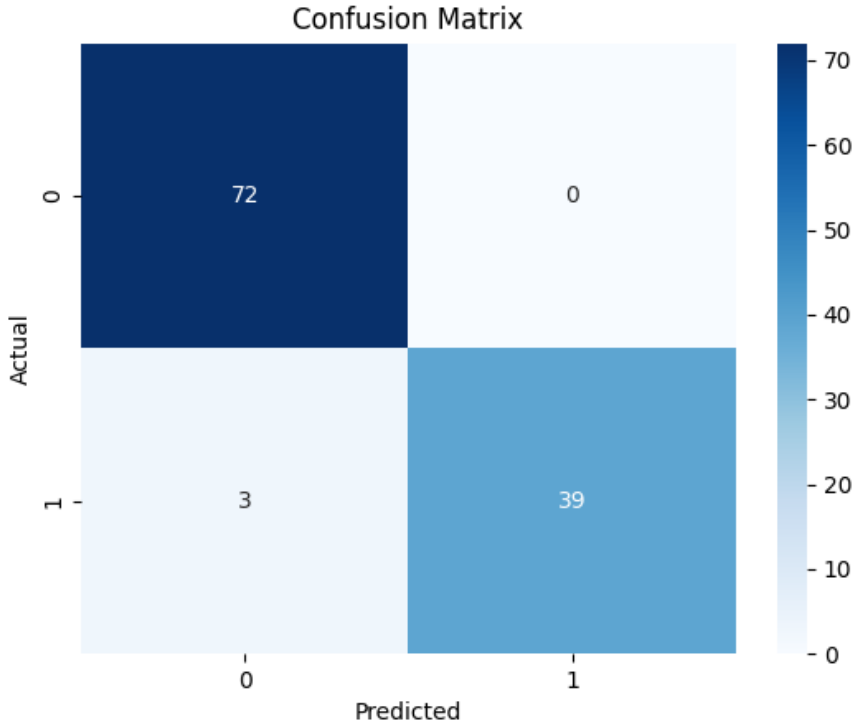

Figure 4.1: Confusion Matrix for Random Forest Classifier

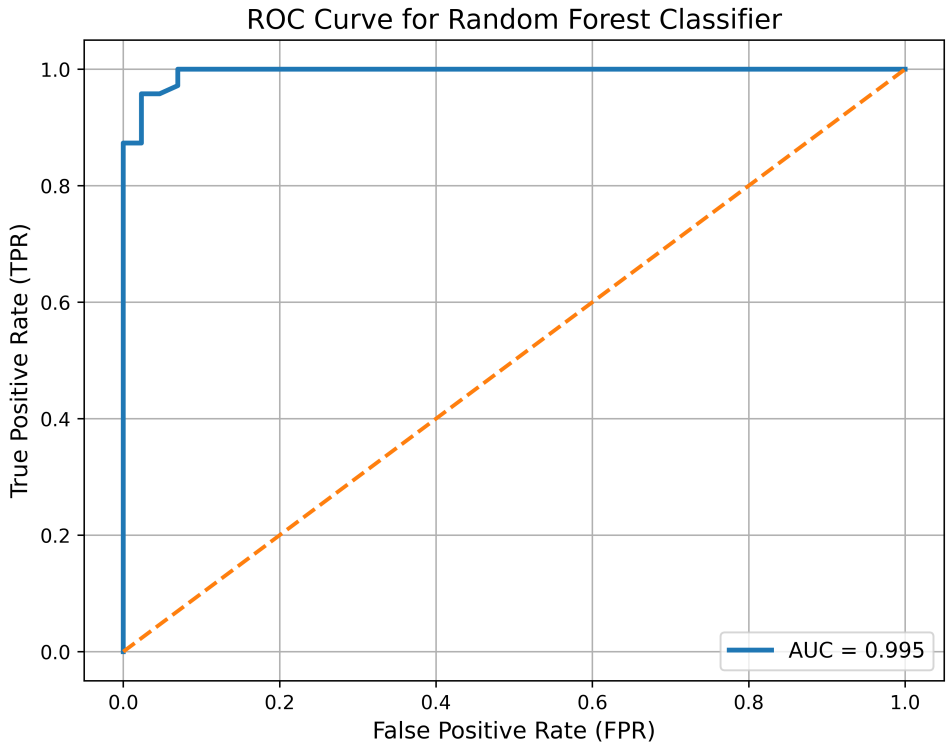

Figure 4.2: ROC Curve for Random Forest Classifier

In general, the ROC analysis confirms that the proposed method maintains a high level of

sensitivity and specificity despite changing the decision threshold.

4.7 SHAP-Based Explainability Analysis

SHAP analysis is performed to assess the global feature importance and just kind of see what is going on with the Random Forest classifier.

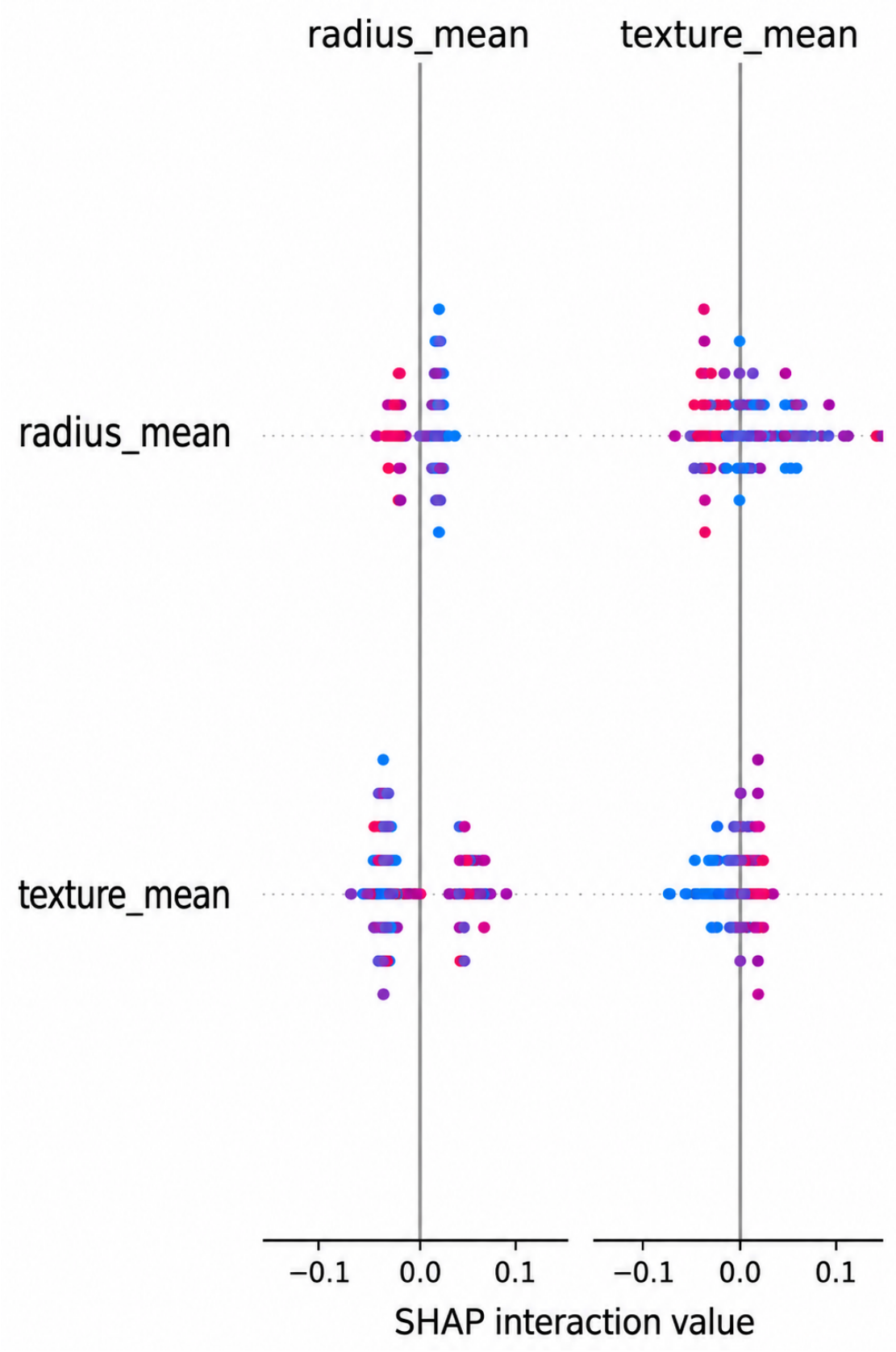

Figure 4.3: SHAP Summary Plot for Global Feature Importance

The SHAP summary plot shown in Figure 4.3 is indicative of the contribution made by vari-

ous features into the model predictions, in general terms. The SHAP summary plot provides a more thorough illustration of the global feature importance along with the direction in which the healthcare-related attributes affect the prediction results. SHAP values are effective when describing the contribution of each feature to the malignant/benign classification process [26].

The closer a particular feature is to the upper part of the SHAP summary plot, the more it is associated with the prediction process. In the suggested method, mean radius, mean texture, worst concavity, mean perimeter, and area mean are highlighted as the more important characteristics used to predict the development of breast cancer.

In this regard, the positive SHAP value shows that a particular characteristic tends to push the model toward malignant tumor prediction, whereas the negative one directs it to classify the breast tumor as benign. Moreover, In SHAP summary plot, Red denotes higher value of feature while blue denotes lower value.

Based on the analysis of the SHAP plot, it may be concluded that the higher values of features such as worst concavity and mean perimeter tend to direct the predictions toward malignant tumors, and the lower values tend to predict benign tumors. The conclusions made coincide with the medical practice in the diagnosis of breast cancer.

Another great benefit of SHAP is that it enables simultaneous local and global interpretability. The global interpretability can be used to understand the overall behavior of the Random Forest classifier, while the local interpretability can be applied to analyze the contribution of each feature on a single patient's prediction individually.

Moreover, SHAP framework also somewhat meets the theoretical requirements of consistency, local accuracy, and missingness [26]. These characteristics make the explanation procedure more robust and trustworthy for the model itself.

Feature importance analysis using SHAP can be helpful for healthcare specialists to understand which medical factors have an effect on the disease diagnosis. Visual explanations can help to verify predictions made by AI models and thus boost confidence in machine learning-assisted decision-making [44].

In comparison to these classic approaches to feature importance, SHAP produces explanations that are more practical and mathematically sensible, despite the fact that one may not necessarily realize it at first glance. Importance scores that do not take into account the actual contribution of the feature in each specific prediction are likely to underestimate the contribution, making the explanation less clear. Contribution scores that are generated by the SHAP approach are calculated according to the principles of cooperative game theory. It means that it does not provide a score of "importance" as "how much it contributed once," but rather "what it contributes in this specific instance." The SHAP summary visualization technique can also reveal some interaction patterns between features and predict trends in the model predictions. It will be easier to understand the inner workings of the algorithm when one knows how it actually makes decisions based on the data.Feature attribution methods have significantly improved interpretability of complex neural and ensemble learning systems [45, 46].

Explainable visualizations obtained using SHAP will make the communication between the AI system and healthcare practitioners more effective. When it is crucial to understand why the model classifies a tumor as malignant or benign, this knowledge is extremely important [16].

4.8 LIME-Based Explainability Analysis

LIME is partially applied in order to provide local explanations for a single prediction produced by the Random Forest classifier for a particular patient.

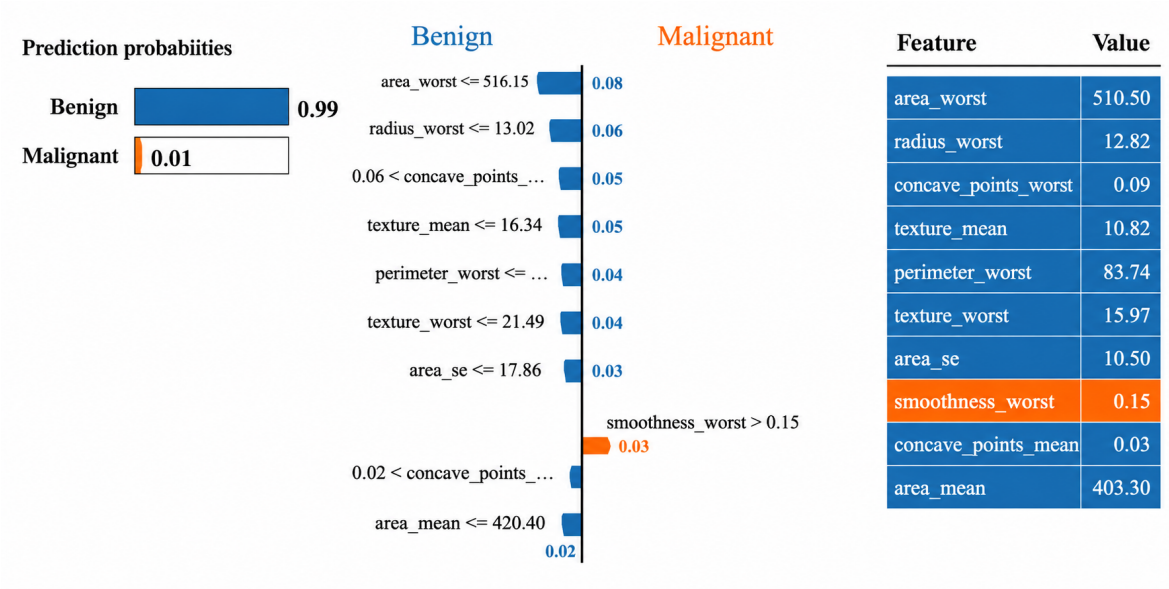

Figure 4.4: LIME Explanation for an Individual Prediction

Figure 4.4 illustrates the LIME explanation of a single patient prediction. From the plot, it becomes obvious which features have contributed to the increase of the outcome value, and those features that have contributed to the opposite effect, i.e., decreased it.

Thus, LIME facilitates the process of gaining an understanding of how exactly the diagnosis was predicted for a particular patient. It means that the level of transparency of AI systems becomes higher, and the confidence in such solutions grows. Local explainability is important in healthcare, because a specific patient interpretation should be provided before making a clinical decision.

All things considered, LIME provides a useful approach to explain AI solutions within healthcare applications. LIME allows interpreting the predictions individually for each patient by determining which features contributed positively or negatively. Local explanations promote transparency and gain clinician confidence.

4.9 Expanded Experimental Results and Interpretation

As a result, the experiment showed that the Random Forest classifier successfully provided accurate prediction results regarding disease classification tasks. Specifically, the classification framework effectively used healthcare features of the Breast Cancer Wisconsin Diagnostic dataset to distinguish between malignant and benign tumors, which was essentially the task pursued. In particular, the preprocessing step played a vital role in improving data quality and boosting the training process. Such aspects as feature scaling and splitting into training and testing sets led to better training conditions and a reduced risk of overfitting. Subsequently, the Random Forest classifier was able to detect more complicated feature dependencies, and generate fairly consistent prediction results while working with the testing samples.

High prediction power indicates that ensemble machine learning algorithms can be utilized for developing healthcare prediction systems with an emphasis on explainability. Quite a lot of samples were assigned to the correct classes, and just a few were assigned to the wrong ones. Importantly, the relatively small number of false negatives could be crucial in health care applications, since an incorrect or late diagnosis could influence patient treatment and clinical decision making in the negative way. In addition, the ROC curve analysis confirmed this conclusion. Indeed, the ROC plot revealed quite good separability between malignant and benign classes. Moreover, at different threshold values, the classifier maintained fairly balanced sensitivity and specificity, indicating its reliability in health care prediction.

Then, the SHAP explainability analysis provided global interpretability of the Random Forest by identifying the most important health care features affecting prediction results. The analysis identified important medical characteristics, such as mean radius, texture, concavity, perimeter, and smoothness, that significantly impacted decisions about disease classification. The SHAP summary plots also demonstrated the contributions of individual features to prediction results. Feature importance analysis played a great role in understanding the model behavior better and allowed healthcare practitioners to identify vital medical features associated with the disease predictions. The explainability approach plays a very important role in health care systems where clinicians usually require some form of interpretation before accepting the automated AI decision-making process, okay. The LIME approach provided local explanations for individual instances. It demonstrated the impact of specific healthcare features on the prediction of an instance. Local explanations added some degree of understanding and helped interpret the rationale behind the classification.

Overall, combining the SHAP and LIME approaches allowed providing both the global perspective and local explanations within the framework of an explainable health care prediction system. SHAP helped gain a general understanding by revealing feature importance at the model level, whereas LIME explained individual cases at the local level. Combining both approaches increased the reliability, interpretability, and transparency of the proposed AI-based health care system.

Compared to the common black box machine learning algorithms, the suggested explainable

framework remains superior in terms of interpretability, even though it does not have any serious impact on the performance of predictive models. Such a combination of accuracy and explainability is particularly important in the context of healthcare, since it requires transparency, accountability, and reliability of algorithms for their practical implementation. Generally speaking, the obtained results demonstrate that the techniques of Explainable Artificial Intelligence have the potential to significantly increase transparency and explainability of healthcare prediction algorithms without affecting their classification capabilities. Overall, the proposed framework shows promise for practical application in trustworthy and explainable AI-based healthcare systems.

4.10 Clinical Significance of the Proposed Framework

Perhaps the greatest advantage for the clinician resulting from the implementation of this framework is improved interpretability of disease predictions. Black box models do not provide a clear chain of logical reasoning leading to their outcomes, which makes it difficult for healthcare professionals to use these predictions in practice. However, the integration of SHAP and LIME techniques allows healthcare providers to understand how the various features in the healthcare dataset affect the prediction outcome.

Moreover, the framework is likely to assist clinicians in early detection and risk evaluation for certain diseases. With an accurate and interpretable prediction system, it becomes easier to identify those patients who might be at a higher risk of developing a particular disease. Early diagnosis matters a lot for example in diseases such as breast cancer, where late detection might, in practice, reduce patient survival rates.

Then there is another clinical advantage which is the improved trust , in AI-assisted healthcare systems. Doctors and other medical practitioners are generally more willing to use machine learning tools when the predictions are transparent and clinically easy to interpret. explainability methods improve confidence in automated healthcare suggestions and they also support evidence-based clinical decisions. The feature importance analysis produced using SHAP is helpful too, because it gives clinical insights about which medical parameters are most influential for disease prediction outcomes.

Healthcare professionals can use these insights to prioritize important clinical indicators while doing diagnosis. LIME-based local explanations add another layer too, because they help with patient specific interpretation by outlining why a model gives a particular result for a specific case. these kind of local explanations can also back personalized healthcare recommendations, and at the same time make it easier for clinicians and patients to talk about AI generated decisions, without that extra mystery .

The explainable framework we proposed may also lower diagnostic uncertainty , and it could assist healthcare institutions in boosting efficiency during medical screening. AI assisted prediction systems can, in practice, reduce the workload for healthcare professionals by giving fast support that is also interpretable. In addition, transparent healthcare AI systems are more ready for real world deployment in clinical settings where accountability reliability

and patient safety are all extremely important. Explainability helps make sure healthcare predictions are understandable auditable and clinically meaningful. Overall , the proposed explainable healthcare prediction framework shows solid potential to support trustworthy, interpretable, and clinically reliable AI assisted healthcare systems.

4.11 Limitations of the Proposed Framework

Although the proposed explainable healthcare prediction framework turned out to be quite effective at forecasting and it did improve interpretability, a few limitations still remain in the present study. The general framework was tested on a structured healthcare dataset which has quantitative clinical features, it is quite simple. But when it comes to real-world, large-scale healthcare datasets, which is heterogeneous and consists of medical features of high dimension, both the performance of framework and the behavior of explanation can be different.

While the Random Forest classifier exhibited good classification performance, ensemble machine learning techniques can require more computation resources when trained on larger data sets. The feature contribution scores computed in the SHAP explainability part can be quite compute-intensive as well, as they're computed many times for many prediction instances.

For LIME, explanations are produced through perturbation driven, local approximations. So, details of the stability of explanations may not be the same each time, even if the setup is similar; a different sample of perturbations and random initializations will shift the results slightly. This could, in turn, influence the consistency in some specific prediction situations.

The current research primarily focuses on supervised machine learning methods and structured tabular data within healthcare settings. Deep learning architectures like convolutional neural networks , and Transformer based healthcare models , were not included within the scope of this work . In addition, the framework was tested in an experimental environment and it was not connected to real-time clinical decision support systems. Practical deployment in healthcare environments may demand more validation, checks for regulatory compliance verification , privacy preservation mechanisms and ethical AI considerations. Still, despite these limitations, the proposed framework managed to show the effectiveness of combining explainable artificial intelligence techniques with machine learning models, in order to enable more transparent and interpretable healthcare prediction systems.

4.12 Comparative Analysis

To evaluate the performance of the proposed framework does its job, the Random Forest classifier is set next to logistic regression, and compared more or less directly.

Table 4.4 shows the comparison outcomes in a summarized kind of way.

Overall, the comparison hints that the explainable version of the framework keeps strong

Table 4.4: Comparative Analysis of Machine Learning Models

Model	Accuracy	F1-Score	Interpretability
Logistic Regression	0.921	0.918	High
Random Forest	0.964	0.963	Moderate
Random Forest + SHAP + LIME	0.964	0.963	High

predictive results , while the interpretability is improved quite a bit.Several researchers have compared post-hoc explanation methods for improving interpretability in machine learning systems [47, 48]. Logistic Regression gives more straightforward explanation, or so it seems, but it ends up with lower accuracy compared to the other choice. Meanwhile, Random Forest brings better classification performance. And when SHAP and LIME are added, the proposed framework ends up offering better transparency , without dropping predictive ability , as it is basically the same accuracy.

4.12.1 Additional Comparative Analysis

In order to verify the effectiveness of the proposed explainable healthcare framework, a comparison analysis is carried out with respect to traditional approaches used in machine learning. The analysis considers several factors including predictive performance, interpretability, transparency, and applicability to healthcare decision support systems.

While traditional models such as Decision Trees and Logistic Regression can be interpreted rather easily, their predictive performance may not be sufficiently high in complex healthcare settings. Ensemble learning techniques such as Random Forest on the other hand, generally exhibit high predictive power but are often referred to as a black box with limited interpretability [28], especially in the clinical domain where clinicians demand explanation.

SHAP technique is used for global feature importance analysis, while LIME provides local explanation of predictions made by the system for individual patients. Such an approach not only makes it possible for clinicians to understand the results of the prediction, but also improves the concept of reliable healthcare AI systems.

Table 4.4 shows a comparative view of several machine learning approaches that are used in healthcare prediction setups.

Table 4.5: Extended comparative analysis of explainable AI models

Model	**Accuracy**	**Interpretability**	**Limitation**
Decision Tree	Moderate	High ,	Lower predictive performance
Logistic Regression	Moderate	High	Limited handling of complex patterns .
Support Vector Machine	High	Low	Difficult to interpret , even in simple setups
Random Forest	High	Moderate	Black-box behavior , not fully traceable
Proposed RF + SHAP + LIME	High	Very High	Increased computational complexity .

The comparative results basically show that the proposed explainable framework gives a better equilibrium between prediction quality and interpretability than most standard machine learning approaches, and well, it still sounds reasonable enough. The explainability methods enhance visibility and support clinicians to interpret the prediction logic more efficiently .
According to Arrieta et al.[49], explainability matters a lot when it comes to building trustworthy Artificial Intelligence systems in high-stakes setups, like healthcare. In a similar way Doshi-Velez and Kim [28] also pointed out that interpret able , and properly accountable machine learning systems are needed if we want them to work in the real world.
In healthcare, when prediction systems are interpretable, the benefits are not only about understanding the model's logic, they also support fairness , responsibility, and overall steadiness in AI-assisted diagnosis. Explainable models allow one to verify if the predictions coincide with actual medical data, and also meet the expectations of clinicians in terms of clinical outcomes. In some cases, this is referred to as practical sense-making.
As demonstrated by Caruana et al.[33], it is possible to have interpretable healthcare models that achieve high predictive capabilities while being completely transparent. Thus, from these findings altogether, it becomes clear that using explainability methods is justified as a part of healthcare AI applications.
From a general perspective, based on the results of the comparative analysis, it can be stated that the proposed approach performs satisfactorily in terms of achieving high levels of predictability along with explainability.

4.13 Discussion

As shown in the experimental results, the use of explainable AI approaches can be effective to increase transparency in machine learning-based healthcare prediction models. The Random Forest classifier is able to consider complex interaction between features in the healthcare dataset, and it performs well. The SHAP approach provides insightful global explanations by identifying important medical features, whereas LIME provides patient-specific explanations of predictions, as if a local perspective. Recent healthcare AI studies have shown that combining predictive machine learning models with explainability frameworks can significantly improve transparency and clinical acceptance of AI-assisted systems [50].

Overall, it can be said that the use of ensemble learning approaches along with explainability approaches increases transparency, trustworthiness and interpretability in AI-based healthcare systems. However, there are some challenges as well. For example, SHAP requires more computing resources than LIME when dealing with large datasets. Meanwhile, LIME explanations may be unstable if the input is perturbed. Furthermore, explainability approaches can introduce additional computational complexity in real-world healthcare applications.

CHAPTER - 5

Conclusion and Future Work

5.1 Introduction

The current chapter serves as the conclusion of the dissertation and includes major results obtained during the research process, along with limitations of the project and further opportunities to investigate related topics. The research was conducted to develop an explainable machine learning algorithm for health care predictions by applying Random Forest SHAP and LIME techniques. Based on the experiments conducted, it could be stated that the use of explainability algorithms in combination with machine learning models contributes to their increased transparency, interpretability, and trustworthiness in the healthcare domain. In other words, an additional explanation provided by the model allows clinicians and other stakeholders.

5.2 Conclusion

The use of machine learning models in the healthcare industry has increased significantly because these models can process large amounts of medical data and produce highly accurate predictions. However, many modern algorithms represent black-box systems where it is impossible to explain the reasons for certain predictions. Therefore, healthcare professionals lack trust in such systems. This dissertation suggested an approach similar to an explainable artificial intelligence framework, partially aimed at improving the interpretability of machine learning models applied in healthcare. Specifically, the framework was developed on the basis of the Random Forest classifier with SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) techniques, whose main goal was to increase interpretability regardless of the complexity of the original model.
The Breast Cancer Wisconsin Diagnostic dataset was selected as a source for experiments on binary classification tasks (malignant/benign tumors). Experimental results showed that the performance of the Random Forest classifier was quite high, which was reflected in the high values of accuracy, precision, recall, and F1-score indicators. Interpretation at the global level was provided by the SHAP technique, and LIME provided interpretation at the local level. Putting the two explainability methods together made it easier to understand how the model behaves, and that also boosted transparency in clinical decision making.
Overall, the proposed framework managed to balance predictive performance with interpretability, so it looks practical for AI-assisted clinical decision-support systems. The re-

sults suggest that explainable AI methods can increase trust, accountability, and reliability in healthcare machine learning applications.

5.3 Research Contributions

The main contributions of this dissertation can be sketched as follows. First , a development of an explainable machine learning framework, for healthcare prediction systems, is proposed. Second , Random Forest classifier has been integrated with SHAP and LIME explainability techniques. Also, there is a comparative study looking at machine learning interpretability alongside explainable AI methods , in practical terms. Then, an experimental evaluation was carried out using the Breast Cancer Wisconsin Diagnostic dataset. Additionally, the analysis covers both global interpretability and local interpretability when healthcare predictions are produced. Overall, the work aims to enhance transparency plus trustworthiness in machine learning based healthcare systems .

5.4 Limitations of the Study

Although the proposed framework managed to show strong predictive performance and decent interpretability, a few limitations still remain. The study used only one structured healthcare dataset and, really, it was mostly about supervised machine learning strategies. SHAP analysis takes more computational power when the dataset becomes big, while LIME explanations can behave a bit differently depending on how perturbations are set. Also, real-time clinical deployment was not explored and deep learning architectures were left out of the experimental analysis entirely.Despite all that, the proposed framework still shows the usefulness of explainable AI methods for healthcare prediction systems.

5.5 Future Scope

Future research could enhance the proposed explainable healthcare prediction framework by incorporating deep learning methods such as CNNs, RNNs, and transformer-based healthcare models. In many cases, these methods have the ability to improve the prediction capability of large-scale, and highly complicated healthcare datasets. However, the current framework mainly relies on structured healthcare data. Multimodal healthcare data such as MRI scans, CT images, histopathological images, genomic data, and even wearable health data may be considered in subsequent frameworks. Apart from that, additional interpretability techniques such as Grad-CAM, Integrated gradients, Partial Dependence Plots (PDP), and counterfactual explanations may be incorporated. These will not only improve interpretability but also increase the visualization capabilities.
Deploying the explainable healthcare system in real-time within clinical decision support environments is another important area for future research. Once these systems get integrated

into hospital information systems, they will become more practical, providing clinicians with valuable diagnostic and treatment planning support. Another important future area will be the computational efficiency of explainability techniques, specifically SHAP and LIME on large-scale healthcare datasets.The computational load can be decreased, and at the same time interpretability can be preserved by applying the previously explained optimization methods and the hybrid explainability approach.
Another area requiring future work is the one of Ethical AI. All future AI healthcare systems will have to be capable to face issues such as fairness,bias detection, accountability, transparency, and privacy preservation in order to guarantee that the system be ethical and responsibly developed in clinical healthcare practices.Future developments in explainable artificial intelligence may include causal reasoning, human-centered explainability and interactive explanation system to enhance real time healthcare decision support [51]. Federated learning and privacy-preserving machine learning may be investigated to achieve health predictions without leaking the privacy of patient personal data. Such methods can considerably help to improve the data security and to allow collaborations between healthcare institutions. In addition, the progress in Explainable Artificial Intelligence seem to hold great promise to increase transparency, trust, clinical relevance, and acceptance of AI healthcare systems. Future work may include the research of multimodal explainability frameworks and privacy-preserving healthcare AI for the medical diagnosis in a secure and transparent way [52, 53]. Recent research has demonstrated that federated and graph based learning model hold promise in distributed and secure AI systems [54]. The use of attention based deep learning models can be helpful for obtaining better performance in explainable AI health care systems [55].

5.6 Final Remarks

Of course the area of Explainable AI (XAI) is extremely important, and has proven to be a necessary ingredient in trust worthy ML models, and for the application areas such as in the domain of healthcare where 'good performance' alone would not be sufficient. However in this proposed system it appears that combining of top performance ML algorithms and the explainability techniques resulted in higher interpretability and an overall similar prediction performance.
Therefore, from an application point of view, this dissertation touches on the broader topic of explainable healthcare artificial intelligence and highlights the significance of interpretable machine learning models, which are necessary for making accountable decisions.

5.7 Social Impact of Explainable Artificial Intelligence in Healthcare

The field of artificial intelligence has greatly improved healthcare processes through better disease diagnosis, predictive analytics, and decision-making. Nonetheless, most traditional

machine learning algorithms are opaque and non-transparent, thereby limiting the ability of healthcare practitioners to have confidence in their results. Explainable Artificial Intelligence (XAI) is a promising solution to overcome this problem as it provides interpretable and clear explanations of machine learning predictions.
The proposed explainable healthcare prediction system using Random Forest, SHAP, and LIME technologies is likely to have a positive social impact by ensuring that explainable AI systems are transparent, fair, and reliable. Explainable AI solutions allow healthcare practitioners to learn about AI prediction reasons and use AI responsibly.

5.7.1 Impact on Healthcare Professionals

Healthcare professionals need transparent reasoning when using AI predictions within their healthcare practice. With SHAP and LIME explainability techniques, healthcare practitioners will be able to understand which medical attributes contribute to the outcome of prediction.
This framework will be helpful for physicians to understand the important medical aspects that contribute to the process of diagnosing diseases. Interpretable AI will improve the confidence of the doctors and provide a basis for evidence-based medicine.

5.7.2 Impact on Patients

Patients tend to distrust automated systems in healthcare because of the black-box nature of the machine learning models. Explainable AI provides better patient satisfaction because it generates explanations that patients can comprehend.
The transparency of healthcare AI systems allows for promoting fairness and accountability in predictions. Patients can analyze how the predictions were generated and, thus, be more confident in using AI in healthcare.

5.7.3 Ethical and Responsible AI

Explainability is crucial when developing ethical and responsible AI systems. In the case of healthcare applications, biased or inaccurate predictions might have harmful consequences for patients. Explainable AI algorithms will detect unfair predictions, feature biases, and inappropriate decision-making patterns in the machine learning systems.
The suggested framework will allow promoting responsible AI approaches by improving transparency, interpretability, and accountability of the healthcare AI systems.

5.7.4 Impact on Society

Explanation-capable AI models for healthcare can increase access to and enhance the quality of medical services, particularly in regions with insufficient healthcare facilities. AI-enabled diagnosis models can help detect diseases early and improve healthcare efficiency.

The inclusion of Explanation-AI in healthcare applications can also decrease diagnostic errors and make medical decisions more consistent. Explanation-based AI models promote adoption of healthcare technologies and improve trust in Artificial Intelligence.
Moreover, explanation-capable healthcare systems can facilitate medical education and research by providing information about relationships between healthcare variables and outcomes of predictions.

References

[1] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[2] A. Esteva, A. Robicquet, B. Ramsundar, and V. Kuleshov, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, no. 1, pp. 24–29, 2019.

[3] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2009.

[4] A. Holzinger, C. Biemann, and C. S. Pattichis, "What do we need to build explainable ai systems for the medical domain?" *arXiv preprint arXiv:1712.09923*, 2017.

[5] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, and S. Tabik, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[6] R. Guidotti, A. Monreale, and S. Ruggieri, "A survey of methods for explaining black box models," *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–42, 2018.

[7] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[8] M. T. Ribeiro, S. Singh, and C. 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*, 2016, pp. 1135–1144.

[9] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[10] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a right to explanation," *AI Magazine*, vol. 38, no. 3, pp. 50–57, 2017.

[11] Z. C. Lipton, "The mythos of model interpretability," *Communications of the ACM*, vol. 61, no. 10, pp. 36–43, 2018.

[12] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, "What clinicians want: Contextualizing explainable machine learning for clinical end use," in *Machine*

Learning for Healthcare Conference. PMLR, 2019, pp. 359–380. [Online]. Available: https://proceedings.mlr.press/v106/tonekaboni20a.html

[13] R. Singh and A. Sharma, "Occluded skeleton-based multi-stream model using part-aware spatial–temporal graph convolutional network for human activity recognition," *Engineering Applications of Artificial Intelligence*, vol. 156, p. 111183, 2025.

[14] A. Arrieta *et al.*, "Trustworthy artificial intelligence for healthcare applications," *Artificial Intelligence in Medicine*, vol. 114, p. 102061, 2021.

[15] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2020.

[16] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *arXiv preprint arXiv:1708.08296*, 2017. [Online]. Available: https://arxiv.org/abs/1708.08296

[17] C. Molnar, *Interpretable Machine Learning*. Lulu.com, 2020. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[18] A. A. Freitas, "Comprehensible classification models: A position paper," *ACM SIGKDD Explorations Newsletter*, vol. 15, no. 1, pp. 1–10, 2014.

[19] S. M. Lundberg, G. Erion, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, 2020.

[20] A. Barredo Arrieta *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[21] D. Gunning and D. Aha, "Darpa's explainable artificial intelligence program," *AI Magazine*, vol. 40, no. 2, pp. 44–58, 2019.

[22] C. Molnar, G. Casalicchio, and B. Bischl, "Generalizable and robust explainable ai in healthcare," *Frontiers in Artificial Intelligence*, vol. 5, p. 863640, 2022.

[23] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[24] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[25] M. T. Ribeiro, S. Singh, and C. 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*. ACM, 2016, pp. 1135–1144.

[26] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[27] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–42, 2018.

[28] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017. [Online]. Available: https://arxiv.org/abs/1702.08608

[29] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[30] Z. C. Lipton, "The mythos of model interpretability," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[31] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Muller, "Explaining deep neural networks and beyond: A review of methods and applications," *Proceedings of the IEEE*, vol. 109, no. 3, pp. 247–278, 2021.

[32] E. F. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, 2020.

[33] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1721–1730, 2015.

[34] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proceedings of the National Academy of Sciences*, vol. 116, no. 44, pp. 22 071–22 080, 2019.

[35] U. Bhatt, A. Xiang, S. Sharma, and A. Weller, "Explainable machine learning in deployment," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 648–657.

[36] G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Information Fusion*, vol. 76, pp. 89–106, 2021.

[37] M. Nauta *et al.*, "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–42, 2023.

[38] Y. Zhang and X. Chen, "Explainable artificial intelligence in healthcare: A review," *Information Sciences*, vol. 570, pp. 547–565, 2021.

[39] S. Ali and F. Khan, "Healthcare explainable ai: Techniques and challenges," *Healthcare Analytics*, vol. 3, p. 100145, 2023.

[40] A. Bussone, S. Stumpf, and D. O'Sullivan, "The role of explanations on trust and reliance in clinical decision support systems," in *2015 International Conference on Healthcare Informatics*. IEEE, 2015, pp. 160–169.

[41] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," *AI Magazine*, vol. 38, no. 3, pp. 50–57, 2017.

[42] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2020.

[43] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[44] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable ai systems for the medical domain?" *arXiv preprint arXiv:1712.09923*, 2017.

[45] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[46] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[47] A. Arrieta *et al.*, "The role of explainability in healthcare ai systems," *AI Communications*, vol. 32, no. 4, pp. 325–340, 2019.

[48] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, 2019.

[49] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[50] A. Holzinger *et al.*, "Explainable ai and multi-modal causability in medicine," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 12, no. 3, p. e1430, 2022.

[51] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[52] R. Miotto *et al.*, "Deep learning for healthcare: Review, opportunities and challenges," *Briefings in Bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.

[53] C. Rudin *et al.*, "Interpretable machine learning: Fundamental principles and 10 grand challenges," *Statistics Surveys*, vol. 15, pp. 1–85, 2021.

[54] A. Samad and V. Bhatia, "Fedgclrec: Federated graph contrastive learning framework for social influence recommendations," *Expert Systems with Applications*, p. 129294, 2025.

[55] R. Singh and A. Sharma, "Stad-convbi-lstm: Spatio-temporal attention-based deep convolutional bi-lstm framework for abnormal activity recognition," *Journal of Visual Communication and Image Representation*, p. 104465, 2025.

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the Thesis Explainable Artificial Intelligence for interpreting machine learning models

Total Pages 48 Name of the Scholar Raviraj

Supervisor (s)

(1) Dr. Abhilasha Sharma

(2) Na

(3) Na

Department Software Engineering

This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software used: Turnitin Similarity Index: 8% , Total Word Count: 14370

Date: 21/05/26

Raviraj
Candidate's Signature

Abhilasha Sharma
Signature of Supervisor(s)

Ravi Raj

updated_ai_report_smallpdf

PlagReportThesis

Document Details

Submission ID

trn:oid:::27535:139863003

Submission Date

May 21, 2026, 3:08 PM GMT+5:30

Download Date

May 21, 2026, 3:24 PM GMT+5:30

File Name

updated_ai_report_smallpdf.pdf

File Size

4.9 MB

48 Pages

14,370 Words

85,003 Characters

8% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 10 words)

Match Groups

- 83 Not Cited or Quoted 8%
  Matches with neither in-text citation nor quotation marks
- 0 Missing Quotations 0%
  Matches that are still very similar to source material
- 0 Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

5% Internet sources
4% Publications
6% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Ravi Raj

updated_ai_report_smallpdf

PlagReportThesis

Document Details

Submission ID

trn:oid:::27535:139863003

Submission Date

May 21, 2026, 3:08 PM GMT+5:30

Download Date

May 21, 2026, 3:24 PM GMT+5:30

File Name

updated_ai_report_smallpdf.pdf

File Size

4.9 MB

48 Pages

14,370 Words

85,003 Characters

0% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

0 AI-generated only 0%
Likely AI-generated text from a large-language model.

0 AI-generated text that was AI-paraphrased 0%
Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

Disclaimer
Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positive results or false negative results), so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.