

ENHANCING BREAST CANCER PREDICTION
WITH RGX MODELS: A TRANSPARENT
APPROACH USING EXPLAINABLE AI

LIANG ZAIYI

UNIVERSITI KEBANGSAAN MALAYSIA

ENHANCING BREAST CANCER PREDICTION WITH RGX MODELS: A
TRANSPARENT APPROACH USING EXPLAINABLE AI

LIANG ZAIYI

PROJECT SUBMITTED IN PARTIAL FULFILMENT FOR THE DEGREE OF
MASTER OF DATA SCIENCE

FACULTY OF INFORMATION SCIENCE AND TECHNOLOGY
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2025

ENHANCING BREAST CANCER PREDICTION WITH RGX MODELS: A
TRANSPARENT APPROACH USING EXPLAINABLE AI

LIANG ZAIYI

PROJEK YANG DIKEMUKAKAN UNTUK MEMENUHI SEBAHAGIAN
DARIPADA SYARAT MEMPEROLEH IJAZAH SARJANA SAINS DATA

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2025

DECLARATION

I hereby declare that the work in this thesis is my own except for quotations and summaries which have been duly acknowledged.

05 February 2025

LIANG ZAIYI
P137059

LIBRARY FETSM

ACKNOWLEDGEMENT

Firstly, I'm very grateful to Malaysia for giving me the opportunity to complete my master's degree.

Secondly, I am very grateful to my supervisor Professor Dr. Zulaiha Ali Othman for guiding me in completing my dissertation. Thank you for her advice and encouragement.

Moreover, I would like to thank all the Professors in the FTSM and data science for their help.

I also want to thank all my family and friends who stayed by my side.

Last but not least, I would like to thank everyone who provided me with encouragement and support, and thank them for giving me the courage to move forward.

LIBRARY FTSM

ABSTRAK

Kanser payudara kekal sebagai cabaran kesihatan global dengan ketepatan rawatan yang sangat terbatas dan penggunaan data yang menyeluruh. Kajian semasa cenderung bergantung kepada set data kecil dan seragam serta kekurangan rangka kerja kukuh dan boleh ditafsirkan, yang mengurangkan kebolehgunaannya secara klinikal. Kajian ini menangani jurang ini dengan meneroka bagaimana kecerdasan buatan yang boleh diterangkan (XAI) dan model pembelajaran mesin hibrid dapat meningkatkan ketepatan dan ketelusan ramalan kanser payudara. Secara khususnya, ia memberi tumpuan kepada empat set data berbeza - Wisconsin, Coimbra, METABRIC, dan SEER - merangkumi pelbagai ciri demografi dan klinikal untuk memastikan kebolehlaksanaan model. Kajian ini mencadangkan rangka kerja ramalan perubatan yang boleh diskalakan dan dipindahkan dengan mengintegrasikan Hutan Rawak (RF), Peningkatan Kecerunan, dan XGBoost ke dalam model integrasi bertindan (RGX) dan menilai algoritma sedia ada. Dalam kajian ini, rangka kerja LIME daripada alat XAI digunakan untuk memberikan penjelasan pada peringkat ciri, yang meningkatkan ketelusan dan kebolehtafsiran model. Aplikasi LIME menyerlahkan ciri utama dalam set data yang berbeza, seperti ciri tumor, tahap gula dalam darah, jangka hayat keseluruhan, dan ciri-ciri pesakit, yang mengesahkan prinsip biologi di sebalik ramalan model. Melalui teknik prapemprosesan dan metrik prestasi yang ketat, termasuk ketepatan dan sensitiviti, model RGX menunjukkan prestasi unggul berbanding model tradisional, mencapai ketepatan dan kekukuhan yang tinggi merentasi pelbagai set data. Kajian ini menangani cabaran utama dalam ramalan kanser payudara, termasuk heterogeniti set data, kebolehskalaan model, dan kekurangan rangka kerja yang boleh ditafsirkan secara menyeluruh. Dengan mengintegrasikan teknologi XAI, kajian ini menyediakan rangka kerja yang boleh diskalakan dan digeneralisasikan untuk ramalan perubatan, di samping mempromosikan keseimbangan antara ketepatan dan kebolehtafsiran model.

ABSTRACT

Breast cancer remains a global health challenge with significant limitations in treatment precision and comprehensive data utilization. Current studies tend to rely on small, uniform data sets and lack a robust interpretable framework, reducing their clinical applicability. This study addresses these gaps by exploring how explainable artificial intelligence (XAI) and hybrid machine learning models can improve the accuracy and transparency of breast cancer predictions. Specifically, it focuses on four different datasets - Wisconsin, Coimbra, METABRIC and SEER - across different demographic and clinical characteristics to ensure the universality of the model. This study proposes a scalable and transferable medical prediction framework that integrates random forest, gradient enhancement, and XGBoost into a stacked integration model (RGX) and evaluates existing algorithms. In this study, LIME framework of XAI tool is used to provide feature-level interpretation, which improves the transparency and interpretability of the model. LIME's application highlights key features of different datasets, such as tumor characteristics, blood sugar levels, overall survival time, and patient characteristics, verifying the biological principles behind model predictions. Through rigorous preprocessing techniques and performance metrics, including accuracy and sensitivity, the RGX model exhibits superior performance compared to traditional models, achieving high precision and robustness over multiple datasets. The study addresses key challenges in breast cancer prediction, including dataset heterogeneity, model scalability, and the lack of a comprehensive interpretable framework. By integrating XAI technology, the study provides a scalable and generalizable framework for medical predictions, promoting a balance between model accuracy and interpretability.

TABLE OF CONTENTS

		Page
DECLARATION		iii
ACKNOWLEDGEMENT		iv
ABSTRAK		v
ABSTRACT		vi
TABLE OF CONTENTS		vii
LIST OF TABLES		x
LIST OF FIGURES		xi
LIST OF ABBREVIATIONS		xii
CHAPTER I	INTRODUCTION	
1.1	Background	1
1.2	Problem Statements	3
1.3	Research Questions	4
1.4	Research Objectives	5
1.5	Research Scope	5
1.6	Significance of Study	6
1.7	Methodology	7
1.8	Project Report Organization	7
CHAPTER II	LITERATURE REVIEW	
2.1	Introduction	9
2.2	Previous Classification Models	9
	2.2.1 Decision Tree	10
	2.2.2 SVM	12
	2.2.3 Naïve Bayes	13
2.3	Previous Hybrid and Ensemble Models	16
	2.3.1 Random Forest	17
	2.3.2 XGBoost	18
	2.3.3 Gradient Boosting	20
	2.3.4 Other Hybrid Models	21
2.4	Application of XAI	23
2.5	Gap in Existing Research	24

2.5.1	Limited Model generalizability across datasets	25
2.5.2	Limited integration of XAI techniques	25
2.6	Summary	26
CHAPTER III METHODOLOGY		
3.1	Introduction	27
3.2	Experimental Design	28
3.3	Data Collection	29
3.3.1	Wisconsin Dataset	29
3.3.2	Coimbra Dataset	30
3.3.3	METABRIC Dataset	31
3.3.4	SEER Dataset	35
3.4	Data Preprocessing	36
3.4.1	Data Cleaning	36
3.4.2	Data Transformation	37
3.4.3	Data Dimensionality	38
3.4.4	Data Standardization	38
3.4.5	Data Visualization	38
3.5	Applied Model and Parameter	39
3.5.1	XGBoost+naïve bayes+voting classifier	39
3.5.2	Stack integration(RGX)	41
3.5.3	Parameter of Models	43
3.6	XAI Explainable	46
3.7	Model Evaluation	47
3.8	Summary	48
CHAPTER IV RESULTS AND DISCUSSION		
4.1	Introduction	49
4.2	Model Results	49
4.2.1	Models	50
4.2.2	Datasets	53
4.2.3	Interpretation of the Findings	67
4.3	Comparison of Previous Literature	69
4.3.1	Model Comparison	69
4.3.2	Comparison of LIME Model Results	70
4.3.3	Conclusion of LIME results across datasets	72
4.4	Summary	73

CHAPTER V	CONCLUSION AND FUTURE WORKS	
5.1	Introduction	75
5.2	Research Summary and Contribution	75
5.3	Limitations	77
	5.3.1 Datasets	77
	5.3.2 Algorithms	78
5.4	Future Works	79
REFERENCES		80

LIBRARY FTSM

LIST OF TABLES

Table No.		Page
Table 2.1	Classification Model Accuracy comparison	15
Table 2.2	Hybrid Model Accuracy comparison	23
Table 3.1	Introduction to attributes in the Wisconsin dataset	30
Table 3.2	Introduction to attributes in the Coimbra dataset	31
Table 3.3	Introduction to attributes in the METABRIC dataset	32
Table 3.4	Introduction to attributes in the SEER dataset	35
Table 3.5	Parameter of models	44
Table 4.1	Results of Wisconsin	58
Table 4.2	Results of Coimbra	62
Table 4.3	Results of METABRIC	65
Table 4.4	Results of SEER	66
Table 4.5	Results of the RGX	67

LIST OF FIGURES

Figure No.		Page
Figure 3.1	Research approach flowchart	28
Figure 3.2	Hybrid model schematic	39
Figure 3.3	RGX schematic	41
Figure 4.1	Wisconsin histogram	55
Figure 4.2	Wisconsin heat map	56
Figure 4.3	Wisconsin pairwise relationship chart	57
Figure 4.4	Coimbra histogram	59
Figure 4.5	Coimbra heat map	60
Figure 4.6	Coimbra pairwise relationship chart	61
Figure 4.7	METABRIC histogram	64
Figure 4.8	SEER histogram	66
Figure 4.9	Comparison with previous studies	70
Figure 4.10	LIME combined with RGX modeling of Wisconsin	70
Figure 4.11	LIME combined with RGX modeling of Coimbra	71
Figure 4.12	LIME combined with RGX modeling of METABRIC	71
Figure 4.13	LIME combined with RGX modeling of SEER	72

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
CNN	Convolutional Neural Networks
DT	Decision Tree
EDA	Exploratory Data Analysis
EMT	Epithelial-Mesenchymal Transition
ER	Estrogen Receptor
GB	Gradient Boosting
HER2	Human Epidermal Growth Factor Receptor 2
IARC	International Agency for Research on Cancer
KNN	k-Nearest Neighbors
LIME	Local Interpretable Model-Agnostic Explanations
LR	Logistic Regression
MAE	Mean Absolute Error
ML	Machine Learning
MLP	Multilayer Perceptron
MSE	Mean Squared Error
NB	Naive Bayes
PSO	Particle Swarm Optimization
RBF	Radial Basis Function
RF	Random Forest
MLP	Multilayer Perceptron
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
SHAP	Shapley Additive Explanations
SVM	Support Vector Machine
WHO	World Health Organization
XAI	Explainable Artificial Intelligence

XGB

XGBoost

LIBRARY FTSM

CHAPTER I

INTRODUCTION

1.1 BACKGROUND

Breast cancer is the most common type of cancer among women in the world and the leading cause of death among women. According to the latest global cancer burden data for 2020 released by the World Health Organization's International Agency for Research on Cancer (IARC), new cases of breast cancer worldwide reached 2.26 million in 2020, surpassing lung cancer (2.21 million cases) for the first time to become the world's largest cancer. In the latest data released by the World Health Organization (WHO 2022), breast cancer still ranks first in the world, with more than 200,000 cases of lung cancer. This makes breast cancer an urgent disease problem that needs to be solved globally and is the focus of this project.

At present, machine learning and data analysis also play a huge role in the field of breast cancer diagnosis and treatment. Among the areas that have been achieved so far, machine learning has done breast cancer screening and diagnosis, pathological analysis, prognostic prediction and risk assessment. Machine learning models such as SVM and Random Forest were developed for ultrasonic image analysis system to automatically detect breast masses (Lee et al. 2020). Companies such as PathAI and Paige.AI have developed AI-powered pathology analysis tools that automatically detect cancer cells in pathology slides, improving the productivity and diagnostic accuracy of pathologists (PathAI. 2022). Memorial Sloan Kettering Cancer Center has developed Breast Cancer Nomogram, an online tool to predict survival and recurrence risk for breast cancer patients by integrating clinical and pathological data from patients (MCKSS.2023). Big data methods have been widely used in the field of breast cancer, covering many aspects from early screening, diagnosis, prognosis prediction,

personalized treatment and so on. These technologies not only improve the accuracy and efficiency of diagnosis and treatment, but also provide a solid foundation for the development of personalized medicine and precision medicine. Machine learning and data analytics have made significant progress in the treatment and prognosis of breast cancer. Through personalized treatment, improved diagnostic accuracy, prognostic prediction, assisted decision-making system and patient management, the treatment effect and quality of life of patients with breast cancer were significantly improved. (Lee et al. 2020) However, there is still a need to further optimize data quality, ensure the fairness and interpretability of models, and enhance the credibility and effectiveness of these techniques through rigorous clinical validation.

However, the existing data of breast cancer are divided into text data and image data. In previous studies, many researchers have made outstanding contributions to text data and image data respectively. The Wisconsin dataset is the most widely used dataset. In the dataset, the researchers used a variety of machine learning and deep learning algorithms to classify the obtained types and find the best classification algorithm. In image data, study (Ji-Yeon et al. 2021) proposes a deep learning-based system for predicting the diagnosis and recurrence of breast cancer. By training and validation utilizing histomorphology imaging data of breast tissue, this work reveals that the prediction accuracy of the system is higher than that of conventional approaches.

XAI has received much attention from bioinformatics and healthcare related researchers. The early XAI systems used SHAP and LIME, which greatly enhanced the interpretability and dependability of advanced machine learning models. For example, Katzman et al. (2016) using SHAP, revealed the relevance of key genetic factors affecting the forecasts of a deep learning model used to project cancer survival. Tonekaboni et al. (2019) found important characteristics such as age, hospitalization history, and specific laboratory test results using SHAP in a study that assessed the risk of readmission of patients through electronic health data. Caruana et al. (2015) predicted the mortality of pneumonia patients by machine learning algorithm and carried out LIME analysis, which also effectively revealed the possible errors in model decision-making, which greatly promoted the improvement of the model and the confidence of clinicians. Early evaluation of these models not only validates the biohealth informatics

performance of XAI technology, but also prepares for the next step of research. Therefore, in this work, XAI technology is also used for data analysis and detection of breast cancer.

Therefore, this paper continues the previous research and makes innovations and breakthroughs in machine learning algorithms for several data sets of breast cancer. By using more complete data sets and updated model algorithms to analyze the previous data sets, the aim is to obtain more accurate indicators and accurate results. At the same time, in order to improve the fairness and interpretability of the model, XAI technology is also used in the study to explain the obtained model. This will facilitate the migration and application of the model in the later stage.

1.2 PROBLEM STATEMENT

Given a high incidence and mortality, Breast cancer has evolved into one of the most important health issues facing women all around in recent years (WHO 2022). Even while management techniques and therapeutic approaches have developed considerably, several important problems still need to be addressed. Among these challenges are the search for accurate early diagnosis, personalizing treatment plans to match individual needs, and filling in the gap of comprehensive, high-dimensional data required for full analysis (Anthis and Kavanaugh, 2020). Moreover, many earlier studies have been restricted by insufficient demographic diversity and small sample counts.

For instance, it is highly sought for in research since the Wisconsin dataset includes clean data free of anomalies or missing values (Khan et al. 2022; Vijaya et al. 2022). However, the limited sample size of the Wisconsin dataset does make it difficult to fine-tune its innovations and accuracy. Due to the extremely small amount of data, the model is difficult to be affected and limited by other factors. Therefore, with additional data, the performance of the model may suffer. Similarly, studies using the METABRIC dataset (Chtouki et al. 2023), while obtaining a larger volume of data, encountered generalization problems because the data came from only one specific place. This means that studies need larger and more complex databases to identify the corresponding influencing factors.

In their studies, researchers frequently use on well-known machine learning methods such Support Vector Machines (SVM), k-Nearest Neighbors (KNN), and Decision Trees (DT) (Amrane et al. 2018). To raise forecast accuracy, ensemble methods and hybrid models have been more underlined in recent years. For more detailed classification, K. Sivakami et al. (2015) suggested a hybrid framework integrating Decision Trees for initial classification with SVM, therefore exploiting the speed of Decision Trees alongside the precision of SVM.

Likewise, Muhammad Umer, Mahum Naveed and their colleagues created a deep learning-based ensemble model (Umer et al. 2022) using Convolutional Neural Networks (CNN) for feature extraction, supplemented by logistic regression and stochastic gradient descent for classification. This method provided notable increases in accuracy. High degrees of accuracy and precision were obtained in another Wadhwa et al. (2023) investigation using a hybrid model including KNN, SVM, and Random Forest (RF). Anggriandi et al. (2023) also utilized a CNN-SVM hybrid model for dermatological diseases, CNN for feature extraction and SVM for classification. The success of these hybrid methods emphasizes their possibility to improve predictive accuracy and precision, which drives this research to investigate their use in raising generalizability and prediction of breast cancer.

However, there are two major unresolved issues in the current experimental research landscape (Sharat et al. 2023). First, breast cancer datasets remain fragmented and separate within the machine learning and data analytics fields. Most studies incorporate at most two datasets of the same type, limiting model generalizability and transferability. Second, while some models perform well on specific datasets, few can effectively interpret the impact of individual features across multiple datasets.

1.3 RESEARCH QUESTION

The main questions of this study are:

1. What improvements in prediction accuracy can the RGX model achieve compared to traditional algorithms on heterogeneous breast cancer datasets?
2. How does the LIME framework enhance clinical trust in the RGX model by

providing feature-level explanations across diverse breast cancer datasets?

1.4 RESEARCH OBJECTIVES

1. To enhance prediction accuracy and robustness for breast cancer using the RGX model on heterogeneous datasets by addressing challenges in data variability and model scalability.
2. To identify key features for breast cancer prediction using LIME-based explanations within the RGX model across diverse datasets, ensuring interpretability and clinical relevance.

1.5 RESEARCH SCOPE

This study aims to develop an interpretable predictive model based on clinical data from breast cancer patients to enhance early diagnosis and treatment precision. Using four diverse datasets—Wisconsin, METABRIC, Coimbra, and SEER—this research included:

1. **Data Collection and Preprocessing:** Ensuring high-quality data by handling missing values, duplicates, and selecting survival-relevant features based on domain knowledge.
2. **Exploratory Data Analysis (EDA):** Utilizing visual and statistical methods to uncover data trends and relationships that inform model building.
3. **Predictive Modeling:** Building and comparing machine learning models to identify the most accurate approach, evaluated on metrics such as accuracy and sensitivity.
4. **Model Interpretability:** Applying LIME from XAI to explain feature contributions, with comparisons to previous best models for precision and generalizability across datasets.
5. **Result Visualization and Conclusion:** Visualizing model performance, highlighting contributions to breast cancer diagnosis and treatment planning.

This research aspires to produce a clinically applicable model that supports personalized patient care and sets a foundation for future refinement and adaptation.

1.6 SIGNIFICANCE OF STUDY

Since precise and timely diagnosis determines the efficacy of treatment, this study is very crucial in solving continuous difficulties in breast cancer prediction. By means of creative machine learning approaches, the study enhances prediction accuracy, therefore lowering diagnostic errors and promoting better informed clinical decision making policies. This precision directly influences patient outcomes and so presents hope for more effective early intervention plans.

This work has been of great help in predicting breast cancer through interpretable artificial intelligence (XAI). Using LIME and other approaches, this work highlights the usability of machine learning models and addresses their opacity. This focus on interpretability enables healthcare professionals to grasp the reasoning behind model predictions, therefore fostering trust and supporting the acceptance of artificial intelligence technology in many different environments. Using artificial intelligence in healthcare is a moral and globally accepted first step in aligning human knowledge with technological advancements.

Furthermore, the study offers priceless insights by means of its thorough investigation of several breast cancer databases. Revealing natural patterns and connections in the data helps one to better grasp the features of diseases, hence guiding next research and clinical procedures. The flexibility of the methodological framework further emphasizes its possible use throughout other cancer types, so providing a scalable and flexible solution in oncological diagnostics.

This study highlights the transforming power of technology in enhancing medical practices and promotes multidisciplinary cooperation between artificial intelligence and the healthcare industry, therefore bridging them. It lays the groundwork for a time when AI-powered technologies will be subtly included into clinical procedures, hence improving the effectiveness and efficiency of patient care results.

1.7 METHODOLOGY

The methodology of this study is divided into four stages. The first stage is to prepare the dataset. The second phase aimed to apply a single machine learning model used in previous studies using four breast cancer datasets, integrate the models and propose a new reference model. The third phase aims to evaluate the models based on the results of the four different datasets and select the best model. The final stage aims to compare the XAI features. A detailed discussion of the phases can be found in Chapter III.

1.8 PROJECT REPORT ORGANIZATION

This thesis is structured into five chapters, summarized as follows:

Chapter I outlines the foundational elements of the study. It includes the background of the research, the statement of the problem, research questions, objectives, scope, and the significance of the study. Additionally, it details the research methodology and provides an overview of the thesis organization.

Chapter II explores the theoretical underpinnings and reviews relevant literature. This chapter delves into the context of machine learning applications in breast cancer prediction, emphasizing the strengths and limitations of prior studies. It highlights how Explainable Artificial Intelligence (XAI) has been utilized and identifies gaps that this research aims to address.

Chapter III focuses on the research methodology, describing the approaches used to achieve the objectives. It presents detailed profiles of the datasets and outlines the deep learning and machine learning models parameter. This chapter also discusses the interpretability methods integrated into the study to enhance model transparency and answer the research questions.

Chapter IV presents the results and findings. It offers a comprehensive analysis of the predictive models' performance, comparing their accuracy, interpretability, and clinical relevance. Visualizations and interpretive insights are provided to facilitate understanding of the results.

Chapter V concludes the study by summarizing the contributions and outcomes. This chapter synthesizes findings from all experiments, highlighting the study's significance in advancing breast cancer diagnostics through AI. It also discusses limitations and proposes directions for future research to build upon the presented work.

LIBRARY FTSM

CHAPTER II

LITERATURE REVIEW

2.1 INTRODUCTION

In recent years, the application of machine learning (ML) in medicine has grown significantly, especially in breast cancer prediction. Many algorithms and models have been proposed in the literature to improve prediction accuracy, but the performance of different algorithms varies on different datasets. In this section, a detailed review of existing studies is presented, ranging from single algorithms and hybrid models to handling unbalanced data and model interpretations, to provide a reference for subsequent studies.

This chapter is structured as follows: Section 2.2 describes the application of a classification machine learning algorithm to breast cancer prediction. Section 2.3 outlines the application of hybrid and integrated models in different areas. Section 2.4 describes the application of Explainable Artificial Intelligence (XAI) to breast cancer prediction. Section 2.5 presents the limitations of existing breast cancer algorithms. Section 2.6 summarizes the chapter.

2.2 PREVIOUS CLASSIFICATION MODELS

The development of machine learning and deep learning has brought biostatistics and health informatics significant changes. Particularly in cancer diagnosis, higher processing capability has enabled researchers to investigate medical data from the 1970s using machine learning methods including logistic regression and discriminant analysis. Early application of pattern recognition technology can effectively interpret the visual data of pathological sections and X-rays, thereby helping to classify tumor types and stages (Koenigkam et al. 2019). These fundamental initiatives allow advanced

machine learning techniques including deep learning and reinforcement learning into the biomedical field to enter the scene. For example, pattern recognition methods were used in 1972 to categorize cancer data, therefore helping doctors to distinguish between several forms of the disease and project future course of development (Kowalski and Bender 1972). Still the cornerstone of contemporary machine learning systems applied in disease diagnosis is this approach.

Using simple machine learning techniques, researchers are analyzing and forecasting medical data across time in breast cancer datasets. Modern studies on breast cancer prediction sometimes stress the need of using certain machine learning algorithms (Rajpoot et al. 2024). Among these, support vector machines (SVM) and K-nearest neighbor models (KNN) are rather often applied. Consistent success of these algorithms in past studies on breast cancer prediction indicates a strong basis for researching more challenging strategies.

2.2.1 Decision tree

a. Model introduction

Decision trees represent a prevalent machine learning technique employed for both classification and regression tasks (Charbuty et al. 2021). This methodology articulates the predictive analysis process via a hierarchical tree structure, systematically partitioning the dataset into increasingly smaller subsets, ultimately culminating in a rule-based decision pathway. Within the context of breast cancer data analysis, decision trees are extensively utilized for various applications, including feature selection, lesion classification, and patient stratification.

1. Information entropy: If the proportion of k-th type samples in the current sample set D is p_k ($k=1,2,3\cdots$), then the information entropy of the sample set is:

$$Ent(D) = - \sum_{k=1}^{|y|} p_k \log_2 p_k \quad (2.1)$$

2. The information gain is:

$$Gain(D, a) = Ent(D) - \sum_{v=1}^V \frac{|D^v|}{|D|} Ent(D^v) \quad (2.2)$$

Generally speaking, the greater the information gain, the greater the "purity improvement" obtained by dividing the data set using feature a . Therefore, information gain can be used to select the attributes to divide the decision tree, that is, to select the attribute with the largest information gain.

b. Previous study

The article explores the application of supervised machine learning algorithms to develop a predictive model for breast cancer (Vijaya, G et al. 2022). Among the methods implemented, decision trees play a pivotal role. Decision trees are hierarchical models that recursively split data into subsets based on feature conditions, ultimately forming a tree-like structure where each leaf node represents a classification outcome. They are widely appreciated for their interpretability and simplicity, making them particularly useful in medical applications where understanding decision pathways is critical.

In this paper, the decision tree technique selected 30 features from the Wisconsin dataset using which to obtain a classification accuracy of 92.39%, therefore separating benign from malignant breast cancer patients. The method consists on separating the data into training and testing groups, creating Gini-based or information gain-based decision trees, and assessing the model's predictive ability. Though they offer some benefits like low data preparation and resistance to outliers, decision trees are prone to overfitting when trees get too complicated. This weakness emphasizes the requirement of optimization techniques including random forests or gradient enhancement or pruning incorporating methods of integration of decision trees.

Although decision trees are good stand-alone models, studies show that integrating other or hybrid techniques could help to raise classification accuracy. The paper did, however, also point up certain shortcomings, including an over-reliance on the Wisconsin dataset and a lack of outside validation. These elements emphasize the need of a more thorough study including several data sets to validate and enlarge the conclusions reached.

2.2.2 SVM

a. Model introduction

SVM is a supervised learning algorithm based on maximization of classification boundaries, which has been widely used in breast cancer data analysis. Its core content is to distinguish benign and malignant tumors by constructing optimal hyperplanes, especially in high-dimensional data and nonlinear classification problems. With the rich clinical features in breast cancer data sets, SVM can not only improve classification accuracy, but also provide important support for early prediction of disease and personalized diagnosis and treatment.

First in SVM is building a hyperplane to divide several data classes from the training set. Within a multidimensional space, a hyperplane is a linear border or surface separating many types of data.

The objective function of SVM can be expressed as :

$$\min = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i \quad (2.3)$$

$$s. t. y_i(wx_i + b) \geq 1 - \xi_i \quad (2.4)$$

$$\xi_i \geq 0, i = 1, 2, 3, \dots, n \quad (2.5)$$

By transforming the objective function by the Lagrange multiplier method, we can obtain the dual form of SVM :

$$\max \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j \gamma_i \gamma_j (x_i * x_j) \quad (2.6)$$

$$s. t. \sum_{i=1}^n \alpha_i \gamma_i = 0 \quad (2.7)$$

$$0 \leq \alpha_i \leq C, i = 1, 2, 3, \dots, n \quad (2.8)$$

b. Previous study

Using the Wisconsin breast cancer data, Vijaya, G, et al.'s 2022 research emphasizes support vector machines (SVM) as the basic method applied to categorize breast cancer cases. Designed to best position data classes with the widest margins, SVM is a supervised learning method building hyperplanes in a multi-dimensional feature space. Three SVM variants: linear, polygon, and radial basis function (RBF) cores are evaluated in this work. Among all the methods, including random forests and decision trees, the linear kernel SVM notably obtained the best accuracy rate of 96.49%.

Data sets can have many features in medical applications since their capacity to handle high-dimensional data sets and create suitable decision boundaries impacts their efficacy, hence SVMs are rather important. In this work, the good performance of linear kernels reveals that linearly separable features characterize breast cancer databases. On the other hand, the radial basis function (RBF) kernel—known for its capacity to control nonlinear interactions—also did really well, therefore stressing its importance in situations when the data consists of more complicated patterns.

Although the research unequivocally points out several shortcomings, support vector machines (SVMs) show pretty good performance. Training SVMs on large datasets is a major difficulty since it is computationally expensive. This is especially valid when applying complex kernel functions such radial basis functions (RBF). Moreover specifically tailored to provide best outcomes are SVMs, sensitive to hyperparameter selection—that is, regularization parameter (C) and kernel-specific settings. The research also stresses the need of greater validation on multiple datasets to examine the generalizability of the conclusions considering the testing limited to the Wisconsin dataset.

2.2.3 Naive Bayes

a. Model introduction

Using a combination of characteristic analysis including mitotic activity, cell size, and mass density, research on breast cancer extensively employs naïve bayes to classify

tumor types. This must-have tool for breast cancer prediction is as its application is in managing noisy or missing data. Early detection capability and diagnosis accuracy both improve with this ability.

Bayesian formula :

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)} \quad (2.9)$$

Known as "prior probability," $P(A)$ is the chance that event A would transpire before event B. $P(A|B)$ is likewise known as "posterior probability," or a reevaluation of the likelihood of occurrence A following event B.

Naive Bayes can decompose the joint probability into the product of the conditional probabilities of each feature :

$$P(X|C) = P(x_1, x_2, \dots, x_n|C) = P(x_1|C) * P(x_2|C) * \dots * P(x_n|C) \quad (2.10)$$

b. Previous study

With the Wisconsin breast cancer dataset, Meriem Amrane et al. (2018) examined how naïve Bayes (NB) and K-nearest neighbor (kNN) might be used to classify breast cancer. This paper highlights that naive Bayes algorithm has shown a quite beneficial performance in probabilistic model simulation decision-making process based on Bayes' theorem. It is consistent over several apps and easy to use since it requires conditional independence of features. Naive Bayesian approaches compute the mean, standard deviation, and probability distribution for every feature and category in the breast cancer classification.

Naive Bayes had a 96.19% accuracy; KNN had a 97.51% accuracy. Still, NB's simplicity and computational efficiency make it a smart decision—particularly for big data sets with high processing needs. Unlike KNN, which depends on distance computations for every fresh prediction, NB's efficiency results from its capacity to prevent training process repetition. Although KNN outperformed NB in this

investigation, the authors observe that NB's scalability and lower processing costs make it remain a competitive rival.

Investigating machine learning methods for breast cancer diagnosis, Yolanda D. Austria et al. (2019) used the Coimbra Breast Cancer Dataset (CBCD). Naive Bayes (NB) is a good model for early classification problems even if its accuracy of 62.38% is lower than that of methods such Gradient Boosting and Support Vector Machines (74.14%). Its simplicity and dependence on probability help to explain this.

These tests show generally that, in terms of simplicity, computational efficiency, and scalability, Naive Bayes offers major benefits even if its accuracy may not always be the best among more sophisticated methods like Gradient Boosting or Support Vector Machine. For datasets with unambiguous distributions and low inter-feature correlations especially, its probabilistic method is quite successful. Future studies should investigate hybrid models like Naive Bayes combined with feature selection methods to improve its predictive accuracy in breast cancer diagnosis.

Table 2.1 summarizes the literature review of four different datasets. The best results of the Wisconsin and SEER datasets were obtained using the classification model, while the best results of the Coimbra and METABRIC datasets were obtained using the hybrid model and are presented in Table 2.2.

Table 2.1 Classification Model Accuracy Comparison

Paper	Dataset	Algorithm	Accuracy
Amrane et al. (2018)	Wisconsin	KNN	0.9928(benchmark)
Vijaya, G et al. (2022)	Wisconsin	SVM	0.96491
Vijaya, G et al. (2022)	Wisconsin	DT	0.9239
Amrane et al. (2018)	Wisconsin	NB	0.961
Austria et al. (2019)	Coimbra	DT	0.6928
Austria et al. (2019)	Coimbra	NB	0.6238
Chtouki et al. (2023)	METABRIC	KNN	0.6875
Chtouki et al. (2023)	METABRIC	SVM	0.7361
Chtouki et al. (2023)	METABRIC	DT	0.6875
Singh et al. (2023)	SEER	KNN	0.6757
Singh et al. (2023)	SEER	SVM	0.7266(benchmark)

2.3 PREVIOUS HYBRID AND ENSEMBLE MODELS

Recent studies show that hybrid models—those which incorporate several machine learning approaches applied to the same breast cancer dataset—tend to outperform single models (Mohammad et al. 2022). Accuracy, precision, recall, and F1 score are just a few of the several evaluation measures where this performance increase is clear. Hybrid models are very successful in capturing complicated patterns in breast cancer data by using the complementing strengths of several algorithms, hence boosting feature representation and detection of minor variations between benign and malignant cases.

Two main advantages of hybrid systems are better generalizing and endurance. Unlike a single model, which could be susceptible to some data patterns or noise, hybrid models reduce overfitting by aggregating a succession of decision processes, hence increasing flexibility to new data (Minskiy and Bober 2022). Combining probabilistic methods such naive Bayes with decision tree-based methods such random forests produces a balanced model that effectively shows feature interactions considering the data probability distribution (Zhang and Wu 2023). Especially in complex datasets such as breast cancer classification, this mix enables more consistent and accurate predictions.

The prospective of hybrid models to enhance breast cancer prediction is investigated in this part. Their exceptional performance makes them appealing candidates for more general therapeutic uses since it offers a strong basis for more research targeted on enhancing treatment relevance and prediction accuracy. Most importantly, hybrid models offer a means to build interpretable, scalable systems fulfilling the high dependability standards in medical diagnostics. Their capacity to mix several algorithms guarantees that they may give clinicians more dependable, accurate, and pragmatic insights, thereby increasing confidence in AI-driven decision-making in healthcare.

2.3.1 Random Forest

a. Model introduction

Random Forest makes advantage of the bagging (Bootstrap Aggregation) method in ensemble learning. Multiple decision trees provide the basis estimators of this homogenous estimator. Every tree learns on a separate subset of the data, and their forecasts are aggregated to generate a result at last.

Breast cancer prediction applications include tumor categorization, feature selection, and condition evaluation make random forest widely employed. It generates several decision trees, each trained on a separate portion of the data, hence improving tumor categorization (Mohammad et al. 2022). This lets it reasonably classify cancers depending on clinical criteria (such as age, tumor size, and texture) and imaging data (such as mammograms and ultrasonic scans). Random Forest is very good in differentiating benign from malignant tumors since it combines the forecasts of several decision trees to provide enhanced accuracy and stability over individual models.

b. Previous study

Mohammad et al. (2022) used Wisconsin breast cancer data to conduct experiments on breast cancer identification. They used RF, KNN, LG and other algorithms for the experiment. The best performance was logistic regression (98%). Random forest (96%) also showed excellent performance. Random forest is an integration of multiple decision trees and has excellent performance in the field of breast cancer detection and identification.

In another study, Khaoula Chtouki et al. (2023) used Random Forest to predict 5-year survival rates in breast cancer patients using METABRIC data. Random Forest achieved a classification accuracy of 75.5%, which was comparable to other models such as Support Vector Machines (74.7%) and Adaptive Boosting (78%). This study examines Random Forest's use of bagging to generate several decision trees, as well as its ability to prioritize feature relevance. This aids in identifying critical survival factors such as tumor size and disease stage.

In an article published by Deepti Singh et al. (2023), random Forest was particularly good at predicting breast cancer. They compare random forests to other machine learning algorithms and mention a new, little-used dataset, the SEER dataset. In the paper, they proposed that random forest has always scored well in Accuracy, Precision and Recall, which is due to the strong integration technology of random Forest.

2.3.2 XGBOOST

a. Model introduction

Many machine learning techniques can identify breast cancer; XGBoost shines because it makes use of "weak learners"—decision trees. Every next tree is built to correct the mistakes of its ancestors. This repeated approach helps the computer to detect complex trends in breast cancer data, hence improving its predicting power.

To start an initial prediction for breast cancer risk, the technique starts with a basic baseline model, say the mean value of tumor characteristics in the training set for regression jobs. It then gauges the difference between expected results and actual clinical labeling using a differentiable loss function, say mean squared error (MSE) (Mijanur et al. 2024). This method iteratively improves the predictions by adding new decision trees, therefore guaranteeing accurate modeling of disease progress or diagnosis.

$$\hat{y}^{(0)} = \frac{1}{n} \sum_{i=1}^n y_i \quad (2.11)$$

$$L(y, \hat{y}) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2.12)$$

b. Previous study

Mijanur et al. (2024) completed the prediction of breast cancer data using the Wisconsin breast cancer dataset on which several different machine learning algorithms were applied. Breast cancer was classed in the experiment using XGBoost. One quite strong hybrid ensemble learning method is XGBoost. To increase the prediction

accuracy and processing performance in data processing, it integrates gradient enhancement with decision tree optimization method.

The XGBoost approach's capacity to manage vast data sets with many features helps to explain one of its primary benefits. In order to concentrate on the most pertinent data points and so lower data noise, this work filtered 21 important features using the Chi-square technique, hence increasing XGBoost's predictive performance. The XGBoost model is shown to have outstanding accuracy and be a useful tool for differentiating benign from malignant patients.

In comparison with other integrated methods, XGBoost algorithm outperforms other model algorithms, especially gradient enhancement (95.32%) and Adaboost (95.91%). This result strongly proves that the XGBoost algorithm is very effective in predicting complex relationships in breast cancer data. When voting classifier was used in combination with SVM, LR, and multi-layer Perceptrons (MLP), this hybrid model achieved an accuracy of 99.42%, surpassing all of the previously listed individual algorithms and integrated methods. This result highlights the important potential of XGBoost to improve prediction accuracy when combined with other methods.

XGBoost's efficiency in this use stems from its parallelizing and tree pruning techniques, which drastically cut calculation time while yet preserving great accuracy. To get best outcomes, the researchers did stress, nevertheless, the need of meticulous hyperparameter adjustment. Less accurate predictions follow from poor performance of the model resulting from suboptimal settings including incorrect learning rates or tree depths. Thus, maximizing XGBoost's potential in breast cancer classification depends on precisely adjusting these hyperparameters.

Although this study reveals that the XGBoost method has major benefits in reaching this prediction objective, its complexity may result in increased computing expenses when used to big data sets. Future studies should investigate integrating the XGBoost algorithm with other modern technologies or approaches to handle this restriction. It is believed that this kind of combination approach will improve the efficiency and performance of the model while so lowering the computational load.

2.3.3 Gradient Boosting

a. Model introduction

On both classification and regression problems, gradient boosting performed admirably. The basic concept of this machine learning method—which gradually adapts weak learners—is that every weak learner concentrates on fixing the mistakes of the one before it (He et al. 2019). Thus, the method is a useful instrument in many different fields, including breast cancer detection since it can efficiently execute weighting and prediction.

Analyzing breast cancer data is where the Gradient Boosting technique shines especially when considering nonlinear correlations and complicated features. Gradient Boosting increases classification and prediction accuracy by iteratively upgrading the model via the mix of several weak learners—such as decision trees (He et al. 2019). For instance, it greatly increases diagnosis accuracy by clearly separating benign from malignant tumors depending on clinical traits including mass size, density, and texture. Furthermore, the method evaluates the relevance of several characteristics, thereby guiding researchers to find important factors in the diagnosis and treatment of breast cancer including age, hormone receptor status, and genetic markers. This capacity to give pertinent features top priority helps to improve model interpretability and enable more focused methods of breast cancer diagnosis.

b. Previous study

Gradient Boosting (GB) is among the primary techniques used for breast cancer prediction on the Wisconsin Breast Cancer Dataset, Kumar et al. found. Together with GB, several boosting methods including XGBoost, AdaBoost, and Stochastic Gradient Boosting (SGB) were tested. In this work, gradient boosting reached a 91.22% accuracy. The research underlines its relevance especially for GB since the performance of hyperparameter tuning depends on elements such the learning rate, number of estimators, and maximum tree depth. Appropriate tuning of these parameters will help to maximize the model and increase the prediction accuracy in activities related to breast cancer classification.

Two major advantages of gradient boosting (GB) are its capacity to use strong loss functions catered to particular problem requirements and its flexibility in managing both regression and classification chores (He et al. 2019). Medical datasets, which commonly have imbalanced classes and noisy characteristics, find this adaptability especially appropriate. Furthermore, GB offers the means for feature priority ranking, therefore enabling researchers to pinpoint, from tumor size, shape, and texture measures, the most important markers of breast cancer (Pinheiro et al. 2024). This feature prioritizing not only increases the accuracy of the model but also offers insightful analysis of the fundamental causes of the diagnosis and prognosis for breast cancer.

2.3.4 Other Hybrid Models

Yasmeen Alslman et al. (2024) document a distributed DDoS (Denial-of-Service) attack detection model by use of a stacked ensemble technique. Random Forest (RF) and eXtreme Gradient Boosting (XGBoost) two robust machine learning methods the model applies to raise the accuracy and efficiency of Intrusion Detection Systems (IDS). By optimizing the complementary strengths of every approach, the authors may improve detection performance. Based on a well-known benchmark in the field of DDoS attack research, the model was tested with the CIC-DDoS2019 dataset therefore providing a comprehensive assessment of the system's possible capacity to identify and reduce such cyber threats.

Using a hybrid approach, Random Forest (RF) and eXtreme Gradient Boosting (XGBoost) are essentially merged in this work both of which are decision-tree-based algorithms. RF offers interpretability and durability by means of its bagging technique, which reduces overfitting and increases model stability (Zhang and Wu 2023). Driven by its boosting method, which iteratively addresses errors generated by past models, XGBoost shows remarkable compute efficiency and accuracy on the other side (Abbasniya et al. 2022). These algorithms are designed in a stacked ensemble architecture whereby the XGBoost forecasts are combined with the RF forecasts via a voting system. This approach uses the features of both models to enhance classification performance and thereby assist the system to detect DDoS attacks with higher dependability and accuracy.

Parizad et al. (2024) showed a hybrid machine learning (ML) model for forecasting energy consumption and electricity price, therefore underlining the need of aggregating models like Random Forest (RF), Gradient Boosting (GB), and XGBoost (XGB). By way of regularization approaches, XGB enhances computing efficiency and precision; RF offers interpretability and robustness against noise; GB catches residual patterns through its sequential learning process using the complementing qualities of each method. Combining these models enables the hybrid system to provide more accurate and consistent forecasts, hence optimizing the forecasting of power pricing and energy demand in dynamic environments.

Emphasizing the interaction of price and demand forecasting inside a hybrid architecture, the work uses XGBoost as a meta-model. This technique improves the general forecast accuracy by including the intermediate results of other models (Parizad et al. 2024). By means of XGBoost as a meta-model, therefore leveraging their strengths, the system effectively aggregates the outputs of several models. Important hyperparameters including learning rate, tree depth, and the number of estimators were also tuned using Particle Swarm Optimization (PSO), so enhancing the predictive accuracy and capacity of the model.

The hybrid model outperformed single techniques in terms of accuracy using considerably lowered RMSE, MAE, and R^2 values. These results present interesting data for prediction activities since they indicate how well hybrid models can manage demanding datasets. The method is rather important in several disciplines, including breast cancer prediction even if the main focus of the work is energy prediction. High-dimensional, nonlinear data provide challenges that hybrid models may effectively manage both times, thereby improving the interpretability and prediction accuracy. Many algorithm integration inside a hybrid framework lets one better control intricate patterns and linkages discovered in the data.

Use a hybrid model RGX based on the above stated study results and talk on the correlation transfer approach of model optimization. The RGX model uses XGBoost as a meta-learner while basic learners combining Random Forest with Gradient Boosting. Combining the benefits of different approaches is supposed to raise the generalizing

ability of the model and prediction performance. The thorough building and experimental findings will be covered in chapter three.

Table 2.2 Hybrid Model Accuracy Comparison

Paper	Dataset	Algorithm	Accuracy
Khan et al.(2022)	Wisconsin	RF	0.96
K.Sivakami et al.(2015)	Wisconsin	DT-SVM	0.91
Wadhwa et al. (2023)	Wisconsin	KNN-SVM-RF	0.9856
Austria et al. (2019)	Coimbra	GB	0.7414(benchmark)
Austria et al. (2019)	Coimbra	RF	0.7031
Chtouki et al. (2023)	METABRIC	Adaboost	0.7870(benchmark)
Chtouki et al. (2023)	METABRIC	RF	0.7778
Singh et al. (2023)	SEER	RF	0.725

2.4 APPLICATION OF XAI

Growing machine learning use in breast cancer prediction in recent years has brought model transparency and interpretability front stage. LIME (Local Interpretable Model-agnostic Explanations) is a common instrument in Explainable AI (XAI) used to grasp the predictions of complicated models (Ribeiro et al. 2018). LIME highlights the significance of several aspects in the breast cancer prediction model by building a local linear model around a given prediction.

LIME, for example, can assist assess how variables such tumor size, texture, smoothness, and other criteria affect the predictions of the model in breast cancer prediction, so offering important information for clinical decision-making (Ribeiro et al. 2018). Improving the dependability of the model helps LIME also help researchers and medical practitioners find important biomarkers, therefore guiding the creation of individualized treatment plans.

Though explainable artificial intelligence (XAI) offers several advantages for breast cancer data analysis, its use is still in early years and few studies in this field. Since XAI technologies are somewhat new and rely on advanced ideas based on sophisticated algorithms and mathematical frameworks, they are largely responsible for this. But as XAI tools and resources expand, demand for processing capabilities and

long training durations using SHAP or LIME approaches also has lately expanded. Notwithstanding these challenges, XAI has great opportunity to improve model transparency and dependability, model robustness, and identify important traits supporting doctors in their diagnosis activities.

Integrated explainable artificial intelligence (XAI) and machine learning approaches for data analysis and study for the Mendeley, Silva et al. (2023) breast cancer dataset. In this study, the authors show how XGBoost in data analysis utilizing SHAP approaches finds significant classifications and key characteristics in datasets with accuracy by means of SHAP approaches. This work emphasizes how XAI may assist to increase model transparency.

Using SHAP to emphasize the most important variables and clear the contributions of hereditary elements in a breast cancer survival model, the study cited as Katzman et al.(2016), found These revelations enabled doctors and researchers to better understand model projections, hence improving the clinical decision-making procedures. Furthermore, SHAP study revealed important genetic and clinical features that dramatically affect survival results, thereby highlighting the need of XAI in feature selection and extraction.

Another such is the work by Tonekaboni et al.(2019), which used LIME to emphasize important elements such patient age and tumor size, thereby clarifying a breast cancer prediction model. LIME not only pointed up these important characteristics but also gave a thorough grasp of their influence on the forecasts of the model. Emphasizing the most important aspects for interpretation, this capacity highlights the relevance of XAI in feature selection. Through analysis of these elements, medical practitioners were able to better grasp the rationale of the model, so improving the therapeutic relevance of machine learning predictions in individualized patient treatment.

2.5 GAP IN EXISTING RESEARCH

Even with great advancement, present models of breast cancer prediction have several limits. These difficulties underline the need of improvements to increase clinical

acceptance and prediction accuracy. The significant gaps are shown here together with ideas for next study paths.

2.5.1 Limited Model generalizability across datasets

Many existing breast cancer prediction models demonstrate suboptimal performance when applied to datasets other than their training data, including variations in imaging modalities, patient demographics, and cancer subtypes.

This limitation highlights a lack of robustness and adaptability, restricting the models' clinical utility in diverse settings.

To solve this problem, in this study we propose a new model that combines cutting-edge methods such as domain adaptation and feature selection to solve this problem, thereby improving the performance of combining multiple data sets.

2.5.2 Limited Integration of XAI techniques

Explainable AI (XAI) methodologies, including SHAP (Shapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), are essential for ensuring transparency in medical predictions but remain underutilized in breast cancer prediction models.

The lack of widespread adoption of XAI complicates clinicians' ability to fully trust and rely on AI-driven diagnostics, as the reasoning behind predictions remains unclear.

Future studies should concentrate especially on adding XAI methods even more into prediction models. Enhancing interpretability not only fosters clinical confidence but also facilitates understandable decision-making—a required process for patient treatment.

By means of resolution of these challenges, next research can enhance the dependability, scalability, and openness of breast cancer prediction models, so augmenting their worth and relevance in many clinical environments.

2.6 SUMMARY

The present literature on the use of machine learning for breast cancer prediction is systematically reviewed in this chapter. It assesses the performance of particular algorithms, underlines the benefits of hybrid models, talks on ways to solve class imbalance, and emphasizes how Explainable AI (XAI) may improve model transparency. Although exposing important data patterns and spotting basic trends depends on single models, hybrid and ensemble techniques increase forecast accuracy by combining the characteristics of many algorithms. Ensuring balanced forecasts depends on addressing class imbalance by means of resampling and synthetic data generation, especially since minority classes commonly correspond with malignant instances.

Moreover, XAI techniques are critically required to boost the confidence and openness of model predictions especially in clinical settings where interpretability controls the decision-making. Still, high computer costs, limited data volume and diversity, and the need of more model openness create challenges. The key focuses of upcoming research should be expanding datasets, developing more resource-efficient algorithms, and progressively combining XAI approaches to generate robust and intelligible models. These advances are needed to close the difference between actual clinical applications and research findings on the prediction of breast cancer.

CHAPTER III

METHODOLOGY

3.1 INTRODUCTION

Covering all phases of the research process, this chapter presents the methodological framework applied to achieve the objectives of the study. Targeting consistent, interpretable breast cancer forecasts, all of this—data preparation, model selection, evaluation approaches, explainability strategies—aims to deliver.

This chapter is arranged below: Section 3.2 addresses model selection, sequential data processing, and evaluation criteria of the experimental design. Section 3.3 covers the data preparation stage covering the processing of four separate datasets: WISCONSIN, COIMBRA, METABRIC, and SEER. Section 3.4 covers the numerous methods of data preparation together with their applicability to different datasets. Section 3.5 presents the used models in this work together with the experimental settings for numerous machine learning models. Introduced in Section 3.6 is the XAI model, which under model interpretation aims to enhance model correctness by means of public and open data testing. Section 3.7 guarantees a whole evaluation of the model by outlining the evaluation criteria and validation strategies used including accuracy, precision, recall, and F1 scores. Finally, Section 3.8 provides a framework for reading the model findings in a clinical environment and a summary supporting the justification for every strategy chosen.

3.2 EXPERIMENTAL DESIGN

The research design that has been followed in this project is provided in this section. By applying these steps, the study objectives have been achieved. The main stages of this research approach that have been taken are shown below:

- Step 1: Preparing Datasets.
- Step 2: Modeling.
- Step 3: Evaluation.
- Step 4: XAI explanation.

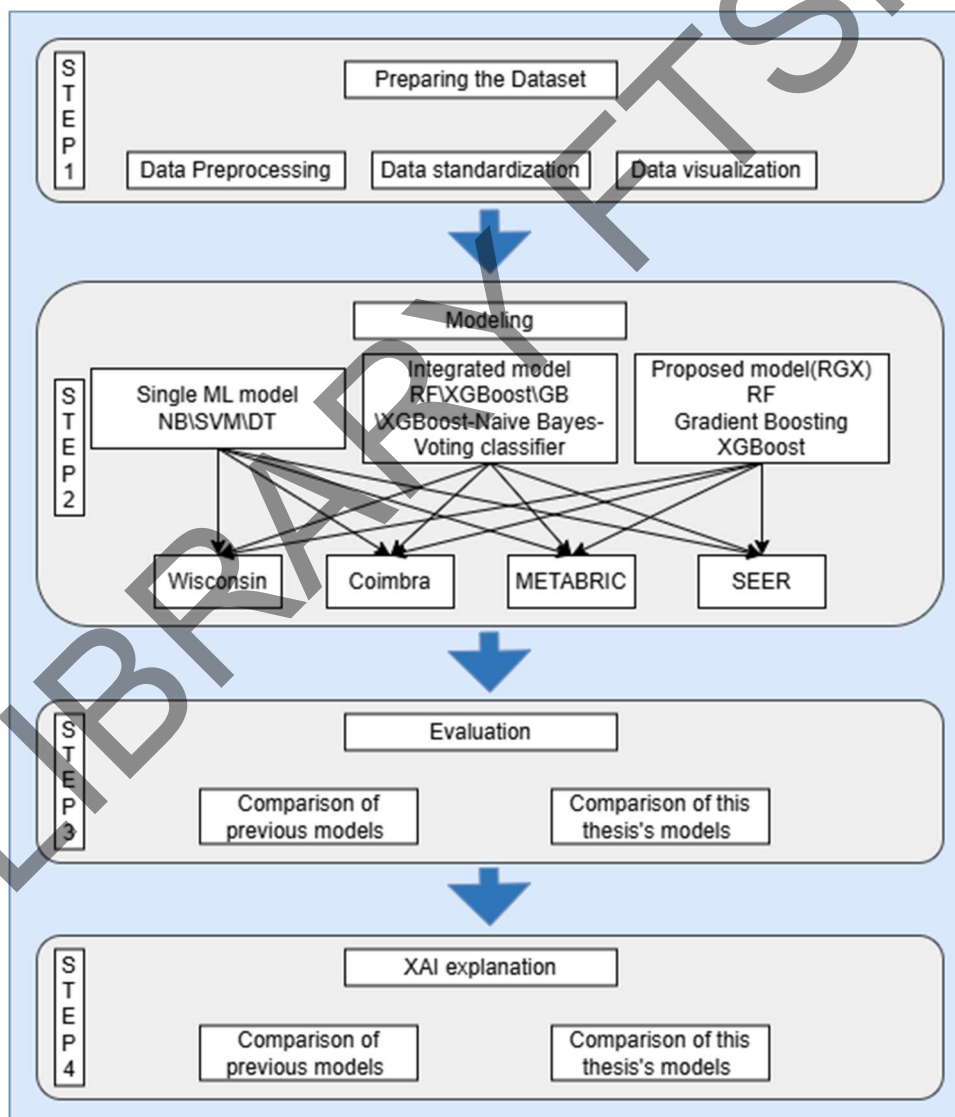


Figure 3.1 Research Approach Flowchart

3.3 PREPARING DATASETS

The aim of Step 1 is to prepare the datasets to input in models. The following paragraphs provide brief information about the dataset. Moreover, the preprocessing for the dataset has also been explained.

In this project, four data sets mentioned in related works. The Wisconsin dataset, Coimbra dataset, METABRIC dataset and SEER dataset are used for processing respectively.

3.3.1 Wisconsin Dataset

The Wisconsin dataset (<https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data>) is obtained from the Kaggle web page. It contains 569 pieces of data and 32 features. The amount of data is relatively small within the usable range. But since most studies use this dataset, we can use this dataset as a reference.

In data mining and machine learning, the Wisconsin dataset is rather famous. For some categorization and prediction issues, its relevance and confidence make it extensively employed in research and practical applications (Ji-Yeon et al. 2021). This dataset mostly reflects the pertinent contents of tumor cell shape, like breast cell size and border. Showing the characteristics of the dataset, digital images of breast nuclei have quite great accuracy and precision. Training effective machine learning models depends on using premium data.

In the Wisconsin dataset, in addition to ID and Diagnosis, 10 different features were explored respectively. At the same time, the average value, standard error, and worst value were calculated for these 10 features, resulting in 30 features. Although the dataset only contains 30 features, they are based on various properties of the cell nucleus such as size, shape, smoothness, etc. These features provide rich information about tumor cells and help the model accurately distinguish between benign and malignant tumors. Table 3.1 shows an introduction to the features of the Wisconsin dataset.

Table 3.1 Introduction to Attributes in the Wisconsin Dataset

No.	Attribute	Type	Description
1	Id	Numeric	Patient code
2	Diagnosis	Numeric	Diagnosis result
3	Radius	Numeric	The radius of the observed nucleus
4	Texture	Numeric	Texture of cell nucleus
5	Perimeter	Numeric	The perimeter of each cell nucleus
6	Area	Numeric	The area of each cell nucleus
7	Smoothness	Numeric	smoothness of cell surface
8	Compactness	Numeric	The compactness of cell morphology. Describes how cells are arranged in space.
9	Concavity	Numeric	an indented area or surface feature on a cell surface
10	Concave points	Numeric	A bulging area or spot on the surface of a cell
11	Symmetry	Numeric	symmetry of cell morphology
12	Fractal_dimension	Numeric	A numerical metric describing the complexity of cell morphology.

3.3.2 Coimbra Dataset

This dataset (<https://www.kaggle.com/datasets/atom1991/breast-cancer-coimbra>) comprehensively explores observed or measured clinical characteristics in 116 breast cancer patients and healthy controls. The Coimbra data set includes 10 features and 116 real data.

Among the several patient clinical and physiological traits found in the Coimbra data set are age, gender, body mass index (BMI), blood pressure, serum biochemical indicators, etc. These characteristics address several facets and help to grasp disease development and prediction holistically. Simultaneously, this dataset has some restrictions on the volume of information. Better generalizing and accuracy are difficult to reach in machine learning and data analysis since the number of data consumed affects the outcomes. This enables researchers to do exact analyses on the dataset without too high computational requirements. With little missing values or outliers, the Coimbra dataset is of rather good quality. This helps researchers to directly use the data for analysis, therefore lowering the demand for intensive preprocessing and data cleaning activities.

This data will be much used by researchers and doctors prepared to probe the complex relationships between clinical features and breast cancer. Including quantitative data helps to identify potential biomarkers linked to breast cancer and offers a whole picture necessary for the creation of predictive models.

Table 3.2 Introduction to Attributes in the Coimbra Dataset

No.	Attribute	Type	Description
1	Age	Numeric	individual's age
2	BMI	Numeric	Body mass index, a measure of body fat based on weight and height
3	Glucose	Numeric	Blood sugar levels, an important metabolic indicator
4	Insulin	Numeric	Insulin levels, a hormone involved in glucose regulation
5	HOMA	Numeric	Homeostatic model assessment, a method to assess insulin resistance and beta-cell function
6	Leptin	Numeric	Leptin levels, a hormone involved in regulating appetite and energy balance
7	Adiponectin	Numeric	Levels of adiponectin, a protein involved in metabolic regulation
8	Resistin	Numeric	Resistin levels, a protein associated with insulin resistance
9	MCP.1	Numeric	Monocyte chemoattractant protein-1, a cytokine involved in inflammation
10	Classification	Numeric	Test results

3.3.3 METABRIC Dataset

This data set (<https://www.kaggle.com/datasets/gunesevitan/breast-cancer-metabric>) comes from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) database. The database, a Canadian-UK project, can help classify breast tumors into additional subtypes with molecular signatures that help determine the optimal course of treatment. Meanwhile, the dataset was collected by research professors at the Cambridge Institute and researchers at the British Columbia Cancer Center in Canada.

In the research of this project, the data set includes 1904 specific data, with a total of 693 columns of features, but there are 520 columns of valid numeric data. In

the study, can choose to exclude these numerical types of data and only rely on some original text and numerical types of data for research. At the same time, some non-numeric data should be selected for data conversion, and it can be entered into a language that the computer can easily recognize.

The Metabric dataset contains clinical data, gene expression data, and other molecular data from thousands of breast cancer patients. This large-scale and multidimensional data allows researchers to conduct deeper analyzes and explore the relationship between molecular and clinical features of breast cancer. At the same time, the data set not only provides clinical information, but also molecular information such as gene expression and gene mutations. This combination of clinical and molecular information allows researchers to study breast cancer from different levels and deeply explore its pathogenesis and therapeutic targets. The Metabric data set covers a variety of breast cancer subtypes and molecular subtypes and has a certain degree of representativeness. This allows researchers to conduct diverse studies on this data set, exploring differences and commonalities between different subtypes. Table 3.3 shows the characteristics of the METABRIC dataset.

Table 3.3 Introduction to Attributes in the METABRIC Dataset

No.	Attribute	Type	Description
1	Patient_id	Object	Patient ID
2	Age_at_diagnosis	Float	Age of the patient at diagnosis time
3	Type_of_breast_surgery	Object	Breast cancer surgery type: 1- MASTECTOMY, which refers to a surgery to remove all breast tissue from a breast as a way to treat or prevent breast cancer. 2- BREAST CONSERVING, which refers to a surgery where only the part of the breast that has cancer is removed
4	Cancer_type	Object	Breast cancer types: 1- Breast Cancer or 2- Breast Sarcoma
5	Cancer_type_detailed	Object	Detailed Breast cancer types: 1- Breast Invasive Ductal Carcinoma 2- Breast Mixed Ductal and Lobular Carcinoma 3- Breast Invasive Lobular Carcinoma 4- Breast Invasive Mixed Mucinous Carcinoma 5- Metaplastic Breast Cancer

to be continued...

...continuation

6	Cellularity	Object	Cancer cellularity post chemotherapy, which refers to the amount of tumor cells in the specimen and their arrangement into clusters
7	Chemotherapy	Int	Whether or not the patient had chemotherapy as a treatment (yes/no)
8	Pam50+_claudin-low_subtype	Object	Pam 50: is a tumor profiling test that helps show whether some estrogen receptor-positive (ER-positive), HER2-negative breast cancers are likely to metastasize (when breast cancer spreads to other organs). The claudin-low breast cancer subtype is defined by gene expression characteristics, most prominently: Low expression of cell-cell adhesion genes, high expression of epithelial-mesenchymal transition (EMT) genes, and stem cell-like/less differentiated gene expression patterns
9	Cohort	Float	Cohort is a group of subjects who share a defining characteristic (It takes a value from 1 to 5)
10	Er_status_measured_by_ihc	Float	To assess if estrogen receptors are expressed on cancer cells by using immune-histochemistry (a dye used in pathology that targets specific antigen, if it is there, it will give a color, if it is not there, the tissue on the slide will be colored) (positive/negative)
11	Er_status	Object	Cancer cells are positive or negative for estrogen receptors
12	Neoplasm_histologic_grade	Int	Determined by pathology by looking the nature of the cells, do they look aggressive or not (It takes a value from 1 to 3)
13	Her2_status_measured_by_snp6	Object	To assess if the cancer positive for HER2 or not by using advance molecular techniques (Type of next generation sequencing)
14	Her2_status	Object	Whether the cancer is positive or negative for HER2
15	Tumor_other_histologic_subtype	Object	Type of the cancer based on microscopic examination of the cancer tissue (It takes a value of 'Ductal/NST', 'Mixed', 'Lobular', 'Tubular/ cribriform', 'Mucinous', 'Medullary', 'Other', 'Metaplastic')
16	Hormone_therapy	Int	Whether or not the patient had hormonal as a treatment (yes/no)

to be continued...

...continuation

17	Inferred_menopausal_state	Object	Whether the patient is post menopausal or not (post/pre)
18	Integrative_cluster	Object	Molecular subtype of the cancer based on some gene expression (It takes a value from '4ER+', '3', '9', '7', '4ER-', '5', '8', '10', '1', '2', '6')
19	Primary_tumor_laterality	Object	Whether it is involving the right breast or the left breast
20	Lymph_nodes_examined_positive	Float	To take samples of the lymph node during the surgery and see if there were involved by the cancer
21	Mutation_count	Float	Number of gene that has relevant mutations
22	Nottingham_prognostic_index	Float	It is used to determine prognosis following surgery for breast cancer. Its value is calculated using three pathological criteria: the size of the tumour; the number of involved lymph nodes; and the grade of the tumour.
23	Oncotree_code	Object	The OncoTree is an open-source ontology that was developed at Memorial Sloan Kettering Cancer Center (MSK) for standardizing cancer type diagnosis from a clinical perspective by assigning each diagnosis a unique OncoTree code.
24	Overall_survival_months	Float	Duration from the time of the intervention to death
25	Overall_survival	Object	Target variable whether the patient is alive or dead.
26	Pr_status	Object	Cancer cells are positive or negative for progesterone receptors
27	Radio_therapy	Int	Whether or not the patient had radio as a treatment (yes/no)
28	3-gene_classifier_subtype	Object	Three Gene classifier subtype It takes a value from 'ER-/HER2-', 'ER+/HER2- High Prolif', nan, 'ER+/HER2- Low Prolif', 'HER2+'
29	Tumor_size	Float	Tumor size measured by imaging techniques
30	Tumor_stage	Float	Stage of the cancer based on the involvement of surrounding structures, lymph nodes and distant spread
31	Death_from_cancer	Int	Whether the patient's death was due to cancer or not (yes/no)

3.3.4 SEER Dataset

The SEER Breast Cancer Dataset (<https://www.kaggle.com/datasets/reihanenamdari/breast-cancer>) is a public database provided by the SEER Program of the National Cancer Institute of the United States for cancer epidemiological research (IEEE, n.d.). The SEER program collects and provides data on cancer incidence, survival, treatment, etc. across the United States. The breast cancer dataset is a part of it, which is mainly used to study and understand the epidemiological characteristics, treatment effects, survival rates, etc. of breast cancer. The SEER dataset mainly includes some patient characteristics, including age, gender, etc. It also includes some tumor characteristics, such as tumor size, grade, histological type, etc. These data are of great significance for studying the pathogenesis, treatment effect, and survival prognosis of breast cancer.

This dataset contains 4025 valid data and 16 features. The dataset includes not only numerical data but also string data. Table 3.4 shows the characteristics of the SEER dataset.

Table 3.4 Introduction to Attributes in the SEER Dataset

No.	Attribute	Type	Description
1	Age	Numeric	Patient age
2	Race	Object	Patient race
3	Marital Status	Object	Get married or not
4	T Stage	Object	T stage, the primary site or size of the tumor
5	N Stage	Object	Lymph node metastasis
6	6th Stage	Object	The sixth installment
7	differentiate	Object	Surgical solution
8	Grade	Numeric	Level now
9	A Stage	Object	The difference in location
10	Tumor Size	Numeric	Size
11	Estrogen Status	Object	Estrogen Status
12	Progesterone Status	Object	Progesterone Status
13	Regional Node Examined	Numeric	Regional Node Examined
14	Regional Node Positive	Numeric	Regional Node Positive
15	Survival Months	Numeric	Date of survival from diagnosis to date
16	Status	Object	Current living status

3.4 DATA PREPROCESSING

Data preprocessing is a crucial step in data analysis and machine learning. Its purpose is to prepare raw data for subsequent analysis and modeling. In this study, we implement appropriate data preprocessing methods for four different data sets according to their respective data characteristics and data types.

For the Wisconsin and Coimbra datasets, we focused on data integrity and outlier handling. This includes detecting and dealing with data issues such as missing values, outliers, etc. that may affect the model results to ensure the quality and reliability of the data. Specific measures include means interpolation, deletion of missing data samples, and outlier detection and processing.

For the METABRIC and SEER datasets, on the other hand, we pay more attention to data conversion and standardized processing so that the data can be more clearly and efficiently entered into the computer for analysis. These conversion measures include the normalization of numerical data, the unique thermal coding of categorical variables and the vectorization representation of text data. These pre-processing steps are designed to improve the operability of the data and the training efficiency of the model, ensuring that all types of data can be fully utilized in machine learning algorithms, thereby improving the performance and predictive power of the model.

3.4.1 Data Cleaning

Data preparation starts among other things with data cleansing. Missing data so have to be found and addressed here at this point. Applied suitable methods for handling missing values—such as eliminating rows with missing data or filling gaps using interpolation and other techniques—are quite crucial since missing data can considerably impair later analysis and modeling. If one intends to guarantee data accuracy and dependability, outliers must also be found and resolved right away using statistical techniques, graphic tools, or specific algorithms. Moreover influencing the results of analysis are repeated records; so, it is imperative to find and eliminate duplicates. Last but not least, the data structure should be correctly changed for more

study and modeling—that is, either changing date and time formats or translating text data to numerical forms.

For the Wisconsin and Coimbra datasets, data preparation tasks began with searches for missing values, duplicate entries, and outliers. Neither of the datasets turned out to have missing or duplicate values. Still, the boxplot revealed oddities. Using the IQR method helped to identify and remove outliers, therefore ensuring data accuracy and clarity. Eliminating points outside the interval of values, the IQR method finds the outlier range. Many kind of data fit this powerful approach. By means of outlier elimination, the dataset is suitable for further research and modeling, therefore providing a robust and consistent base for upcoming projects.

Furthermore used were techniques for missing and duplicate value checks in the METABRIC and SEER databases. Moreover, encoding is needed since these datasets include non-numeric data. This work efficiently codes both datasets using a thermal coding technique, therefore permitting simpler processing for the following models.

3.4.2 Data Transformation

Data transformation is another important step in data preprocessing. At this stage, various transformation operations need to be performed on the data to make it more suitable for analysis and modeling. Feature selection is a critical step that involves selecting feature columns that are meaningful for analysis or modeling. This can be selected based on domain knowledge, feature importance assessment, etc. Secondly, feature extraction is to extract new features from the original data to enhance the representation ability of the data. For example, extract keywords from text, or extract feature vectors from images, etc.

In the research, for the Wisconsin and Coimbra datasets, missing value processing, outlier detection and processing, and feature selection are required. Correlation analysis, PCA and other methods can be used to screen out the most influential features of the model prediction. The processing of the unique thermal coding in the METABRIC and SEER data sets are very convenient for the processing

of machine learning algorithms. Therefore, in these four data sets, we have carried out partial processing of data characteristics.

3.4.3 Data Dimensionality

Data preparation requires first a reduction in data dimensionality. At this point, the dimensions of the data have to be shortened by conserving the most important information and therefore minimising complexity. Common methods consist in feature selection and main component analysis. While also improving the efficiency and performance of later analysis and modeling, dimensionality reduction helps lower storage needs and computing complexity.

3.4.4 Data Standardization

One important phase of preparation is data normalizing. The data should be adjusted to have zero mean and unit variance in line with the criteria of the model. Standardizing helps to lower bias resulting from varying size across features, therefore strengthening the model's dependability. Data standardizing and normalizing methods are applied in the Wisconsin and Coimbra datasets to bring all features to the same scale, therefore enabling simpler processing in later stages.

3.4.5 Data Visualization

Data visualization is quite important for preparedness. In exploratory data analysis, visualizing tools help scientists to better understand linkages, data distribution, and patterns. For upcoming research and modeling, this method exposes trends and latent insights that could be really beneficial.

In summary, data preprocessing is a key step in data science, providing a reliable foundation for subsequent data analysis and modeling. In the four data sets proposed, we adopt the methods suitable for each data set for data preprocessing, and have different emphases for numerical data and non-numerical data. Through the steps of data cleaning, transformation, dimensionality reduction, standardization and visualization, the quality and availability of data can be guaranteed, thus improving the

accuracy and efficiency of analytical modeling. Therefore, without data preprocessing, follow-up work may be meaningless.

3.5 APPLIED MODEL AND PARAMETER

The hyperparameters of the models used in this study were selected to ensure optimal performance. The selected values for models were determined through grid search and cross-validation techniques. Table followed presents the key parameters for each model, along with their respective values and descriptions.

3.5.1 XGBoost+naïve bayes+voting classifier

The combination of XGBoost, naive Bayes and voting classifiers provides a robust ensemble learning framework for breast cancer prediction. Figure 3.2 is a flowchart combining these three models.

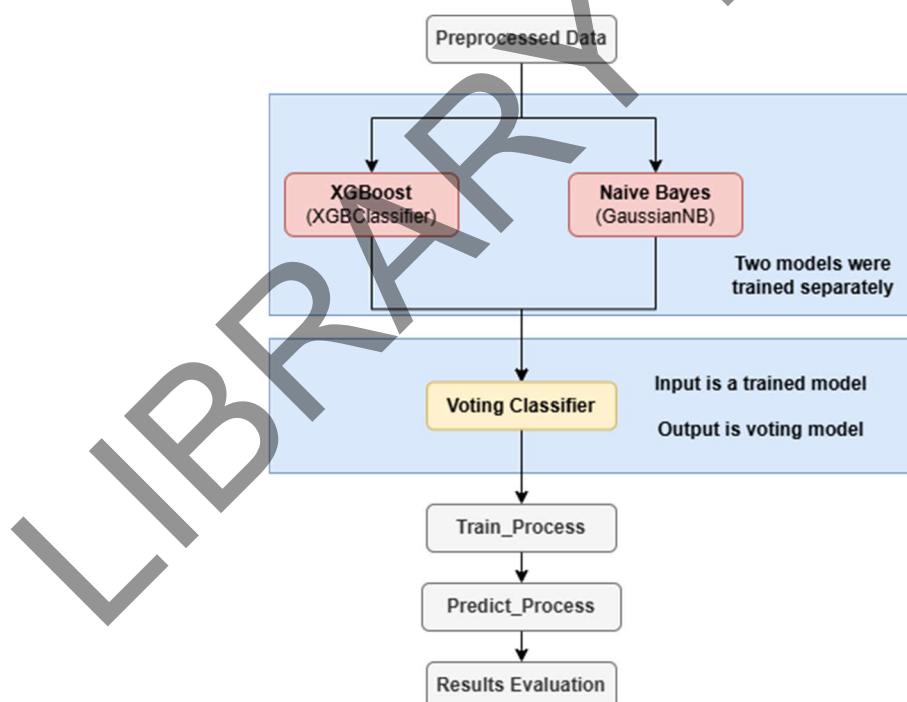


Figure 3.2 Hybrid Model Schematic

The following code is python code that shows how the model is combined in the figure above:

```
#Initialize the base model

xgb_model = XGBClassifier(n_estimators=100, random_state=42)
nb_model = GaussianNB()

# Inflow hard voting

voting_model = VotingClassifier(estimators=[('xgb', xgb_model),
('nb', nb_model)], voting='hard')
```

In this model, an ensemble learning approach was adopted, and a Voting Classifier was used to perform the classification task by combining XGBoost model and Naive Bayes model. First, the data set is split into a training set and a test set, with 80% used for training and 20% for testing. Next, two base models are initialized: the XGBoost model and the Naive Bayes model. The XGBoost model is initialized by XGBClassifier and trained by training set X_{train} and target variable y_{train} . Naive Bayes models are initialized using GaussianNB and trained using the same training data. After the training is completed, the two base models get their own training models.

Next, the Voting Classifier is used to combine the XGBoost model with the naive Bayes model to achieve the final classification prediction. By hard voting on the predicted results of the two base models, the voting classifier determines the final classification result of each sample according to the majority voting principle. In this process, XGBoost and Naive Bayes models predict each sample separately, and the voting classifier selects the category with the highest number of predictions in the two models as the final prediction for that sample.

Voting Classifier aggregates the predictions of XGBoost, Naive Bayes, and potentially other models, to deliver a final decision. By employing a hard voting reduces individual model biases and improves overall classification stability.

Finally, the voting classifier makes predictions on the test set X_{test} , generates the prediction result y_{pred} , and compares it with the actual label y_{test} to evaluate the performance of the model. To fully measure the effectiveness of the classification

model, we used common evaluation metrics such as accuracy, accuracy, recall, and F1 scores.

3.5.2 Stack integration(RGX)

In this paper, we apply a stacked integration model. This model is based on Random Forest and Gradient Boosting, and uses XGBoost as the initial metamodel of this model. Using these models, we can effectively get good results on appropriate data sets, which is conducive to achieving the generalization ability and robustness of the model. In this study, we call the newly proposed model RGX integration model. Figure 3.3 is a flow diagram of the RGX model combination.

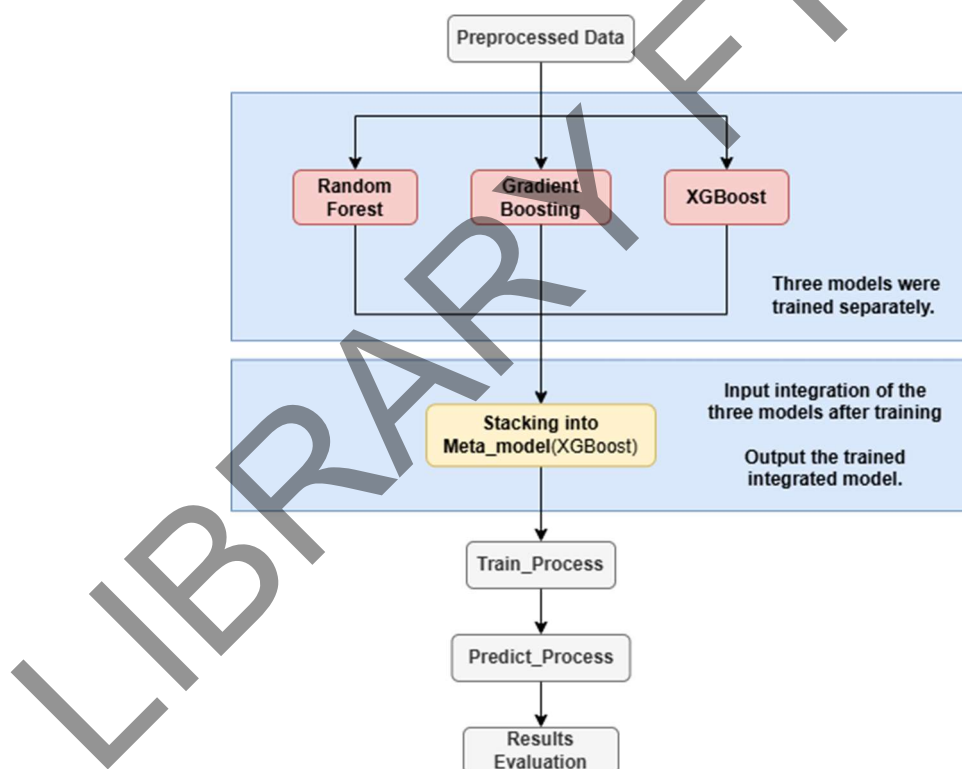


Figure 3.3 RGX Schematic

The following code is python code that shows how the model is combined in the figure above:

```
# Initialize the metamodel (XGBoost)
meta_model = XGBClassifier()

# Build a stacked integration model
estimators = [
    ('rf', rf_model),
    ('gb', gb_model),
    ('xgb', xgb_model)
]
stacking_model = StackingClassifier(estimators=estimators,
final_estimator=meta_model, cv=10)

# Training model
stacking_model.fit(X_train, y_train)
```

The first base model, Random Forest, is initialized using `RandomForestClassifier` and trained on the training set. This model helps in capturing non-linear patterns and reducing overfitting. Similarly, the second base model, Gradient Boosting, is initialized using `Gradient Boosting Classifier` and trained on the same dataset. Gradient Boosting is known for its ability to handle complex non-linear relationships and improve the fit of the model. The third base model, XGBoost, is initialized using `XGBClassifier`, which is also trained on the training data and is renowned for its ability to handle large-scale data and prevent overfitting. Each model is trained independently on the same training set, resulting in three separate trained models.

After training, the three base models—Random Forest, Gradient Boosting, and XGBoost—each make predictions on the data. The stacking classifier then takes these predictions as input and passes them to the meta-model. The meta-model is typically a simpler model, and in this case, another XGBoost classifier is chosen as the meta-model. The meta-model's task is to learn how to combine the predictions from each base model to make the final decision.

Specifically, the meta-model will make a decision based on the predictions from the base models. For example, for the same sample, if Random Forest predicts 0,

Gradient Boosting predicts 1, and XGBoost predicts 0, the meta-model will use these inputs with their corresponding weights to make the final prediction, thereby improving overall classification performance. In this way, the stacking classifier achieves higher accuracy than any single model by leveraging the "wisdom of crowds."

Finally, the Stacking Classifier makes predictions on the test set X_{test} . The predicted results, y_{pred} , are compared with the actual labels, y_{test} , to evaluate the model's performance. To measure the effectiveness of the classification model, standard evaluation metrics such as accuracy, precision, recall, and F1 score are calculated. These metrics provide a comprehensive understanding of how well the ensemble model generalizes to unseen data.

3.5.3 Parameter of Models

The table 3.5 presents the key parameters used for each machine learning model in this study, including the parameters for Decision Tree, Support Vector Machine (SVM), Naïve Bayes, and other models. These parameters control various aspects of the models, such as the criterion for splitting, regularization strength, tree depth, and the number of estimators. The selected values represent typical default configurations that balance performance and efficiency in classification tasks.

Table 3.5 Parameter of Models

Model	Parameter	Value	Description
Decision Tree	Criterion	'gini'	Specifies the function to measure the quality of a split (Gini impurity).
	splitter	'best'	Strategy used to split at each node (chooses the best split).
	max_depth	None	The maximum depth of the tree (grows until all leaf nodes are pure).
	min_samples_split	2	The minimum number of samples required to split an internal node.
SVM	min_samples_leaf	1	The minimum number of samples required to be at a leaf node.
	Regularization	C=1.0	Regularization parameter, balancing margin maximization and classification error.
	Random_state	24	Seed to control random number generation.
Naïve Bayes	kernel	'linear'	Specifies the kernel type to be used in the algorithm.
	priors	None	Model automatically calculates prior probabilities based on training data.
Random Forest	var_smoothing	1e-9	A small value added to variance for numerical stability.
	n_estimators	100	The number of trees in the forest (more trees usually improve performance).
	criterion	'gini'	Specifies the function to measure the quality of a split (Gini impurity).
	max_depth	None	The maximum depth of the tree (grows until stopping criteria are met).
XGBoost	min_samples_split	2	The minimum number of samples required to split an internal node.
	n_estimators	100	The number of trees in the forest (default is 100).
	learning_rate	1	The learning rate used to scale the contribution of each tree (default is 1).
	max_depth	6	The maximum depth of the tree (default is 6, controls overfitting).
	reg_alpha	0	L1 regularization weight (default is 0). L1 regularization weight (default is 0). to be continued...

...continuation

reg_lambda	1	L2 regularization weight (default is 1).
objective	'binary:logistic'	Specifies the optimization objective for binary classification problems.
n_estimators	100	The number of trees in the boosting process (default is 100).
random_state	42	The seed used for controlling the randomness in the training process and data splitting.
test_size	0.2	The proportion of the data to be used as the test set.

3.6 EXPLAINABLE ARTIFICIAL INTELLIGENCE

Explainable AI, also known as XAI, is an intelligent system or tool that integrates multiple AI techniques. It aims to improve the transparency and understandability of artificial intelligence systems, so that people can better understand the decision-making process and principles of AI. With the wide application of AI technology, XAI has become an important field of attention. It not only helps build trust in AI but also addresses issues such as AI ethics and bias. XAI plays a crucial role in the responsible development of artificial intelligence.

Practical experiences with XAI include in human assessment methods to evaluate AI's adaptability, visualization tools to understand deep learning models, XAI decision-making process visualizing methodologies, and decision tree analysis to grasp model decisions. These case studies and practical experiences give readers insightful analysis and direction to help them to grasp and apply Explainable AI.

This work applied the XAI LIME model. Designed to interpret machine learning model predictions, LIME (Local Interpretable Model-agnostics) is a method in Explainable AI. By building local linear models, LIME approximates sophisticated black-box models and helps people to understand why a model generates a given forecast. It generates local, interpretable approximative models to help to clarify the predictions of complex and opaque models, such deep learning or ensemble models.

In this experiment, LIME was used to interpret both the proposed new model and the model with the highest accuracy from previous research. The proposed model is a stacked ensemble model that uses random forest and gradient boosting as base models, with XGBoost as the meta-model. LIME leverages its local interpretability to identify the features most influential in different datasets. It can select specific test cases from the dataset to explain, providing insight into how individual predictions are made. Many new samples are generated in the vicinity of this instance by adding slight perturbations to the original features. At the same time, the proposed stacked integration model is used to predict these adjacent samples, and the prediction probability of each sample is recorded. On the basis of these adjacent samples and their predictions, a simple linear model is trained that approximates the stacked integrated model only

locally. By analyzing the coefficient or feature importance of the local model, the reason for the prediction of the stacked integrated model on this particular instance is explained.

3.7 MODEL EVALUATION

After completing data modeling, it is crucial to compare the resulting models. Common methods of comparison include looking at a model's confusion matrix and accuracy. The confusion matrix provides the model's predictive performance on each category. Including the number of true positives, true negatives, false positives, and false negatives. Through the confusion matrix, you can intuitively understand the performance of the model on different categories, and various evaluation indicators can be calculated, such as precision, recall, and F1 score.

By comparing the confusion matrices and accuracy of different models, you can find the optimal algorithm for your project data set. Typically, the optimal algorithm have higher accuracy across categories and show fewer misclassifications in the confusion matrix. However, the final choice of the optimal algorithm may also depend on the specific needs and optimization goals of the project, such as whether to focus more on accurate predictions of certain categories or the degree of punishment for misclassification.

In order to assess the contribution of various aspects to the prediction outcomes in this work, LIME technology has been used accordingly to the proposed model and the model with the best accuracy in past works. Analyzing the variations in feature contributions between models helps us to understand how features affect model performance and decision-making. This method not only clarifies the internal dynamics of the model but also stresses the importance of features across several model designs, therefore providing useful assistance for additional optimization and enhancement of the predictive model.

In summary, by comparing the confusion matrix and accuracy of the model, the optimal algorithm suitable for the project data set can be found, and further optimization and adjustment can be made according to needs.

3.8 SUMMARY

This chapter defines a comprehensive methodological framework intended to improve breast cancer prediction and interpretation by use of machine learning and Explainable Artificial Intelligence (XAI) approaches. Data preparation, model building, evaluation, and interpretability analysis round out the main processes. Datasets from Wisconsin, Coimbra, METABRIC, and SEER were carefully processed using normalisation, outlier identification, and encoding methods catered to their particular characteristics, therefore guaranteeing high-quality data intake for later study.

The work applied ensemble techniques incorporating multiple machine learning algorithms—including Support Vector Machines (SVM), Random Forests, and Naïve Bayes—in order to attain strong projected accuracy. Furthermore stressed as a technique of evaluating feature contributions and improving model decision-making openness was integration of LIME as an interpretability tool. Evaluation criteria including accuracy, precision, recall, and F1-score were utilized to evaluate model performance with programming done in Python guaranteeing efficiency and adaptability in data administration and presentation.

By articulating each methodological decision and its underlying rationale, this chapter establishes a solid foundation for interpreting results in the following chapters, effectively linking technical advancements with clinical relevance.

CHAPTER IV

RESULTS AND DISCUSSION

4.1 INTRODUCTION

This chapter presents the experimental results obtained by using eight machine learning techniques over four breast cancer datasets. These models were methodically matched to find the most successful one. Evaluating the interpretability of the Explainable Artificial Intelligence (XAI) models then produced the best results. The utilities are rapidly reforming all around the globe in the last years.

Section 4.2 looks at the outcomes the models show during the experiments, therefore laying groundwork for the later comparison study of data. Together with a comparison of the interpretability results from the XAI, the suggested model in Section 4.3 is contrasted with the best-performing model from past studies. At last, Section 4.4 summarizes the main conclusions of the chapter, so clarifying the results of the experiments and the inter-model comparisons, so enabling the identification of the most successful model and the improved experimental outcomes.

4.2 MODEL RESULTS

The four datasets—Wisconsin, Coimbra, METABRIC, and SEER—were evaluated using a series of largely similar models in this study. According to the experimental results, the Wisconsin dataset performed the best. This may be due to its smaller size and lack of redundant data, allowing the model to learn effective features more easily. In contrast, the other three datasets did not perform as well as the Wisconsin dataset, which may be attributed to the inherent complexity of the data. Specifically, the Coimbra, METABRIC, and SEER datasets contain additional features and potential

noise, making the model training process more challenging. Additionally, in the METABRIC and SEER datasets, some categorical features were processed using one-hot encoding. While this technique effectively converts categorical variables into numerical form, it can significantly increase data dimensionality in some cases, reducing data sparsity and potentially impacting the performance of certain algorithms, leading to a slight decrease in accuracy. Nevertheless, the overall experimental results remain highly reliable.

To enhance the robustness of model evaluation and mitigate the potential biases introduced by data partitioning, this study employed the ten-fold cross-validation (10-fold CV) method. Specifically, each dataset was divided into ten equal-sized subsets. In each iteration, nine subsets were used for training the model, while the remaining one subset was used for testing. This process was repeated ten times, ensuring that each subset appears in the test set once. This approach ensures that every sample is adequately trained and tested, avoiding evaluation bias that could arise from uneven data partitioning or the exclusion of certain samples from testing.

In this study, due to the differing distribution characteristics of the Wisconsin, Coimbra, METABRIC, and SEER datasets, the ten-fold cross-validation provides a more comprehensive and reliable assessment of model performance across different data splits. This method not only evaluates the model's accuracy but also calculates the standard deviation of each evaluation, offering a better reflection of the model's stability and generalization ability across different datasets.

4.2.1 Models

In the study, some basic machine learning algorithms were adopted, and some integrated algorithms and hybrid algorithms were also used.

a. Decision Tree

The Decision Tree model showed good results in all four datasets. Similarly, the indicators in the Wisconsin dataset were the highest. In the METABRIC dataset and SEER dataset, one-hot encoding increased the dimension and sparsity of the data,

making the decision tree algorithm face more segmentation choices in the process of building the tree. Therefore, it is easy to produce overfitting or underfitting problems, which affects the classification effect. In order to improve the performance of the decision tree algorithm on complex datasets, pruning techniques and ensemble methods (such as random forests) can be considered to increase the generalization ability and robustness of the model.

b. SVM

SVM beats the models mentioned above. The best results emerged from the Coimbra dataset. Choosing appropriate kernel functions (such as linear kernel, RBF kernel, etc.) would help the SVM algorithm to create a high-dimensional feature space in this dataset, so facilitating good classification. One-hot encoding thereby increases the dimension of the data in the SEER and METABRIC databases. This leads to a spike in SVM's computing cost when deciding the perfect hyperplane in high-dimensional space, therefore affecting the classification impact. Investigating feature selection and dimensionality reduction techniques helps one improve SVM performance on demanding data. By improving kernel functions and hyperparameters and using appropriate regularizing approaches, increase the generalizing capacity of the model. Linear SVM or kernel approximation methods could potentially help you to reduce computational complexity while managing high-dimensional data.

c. Naïve Bayes

The size and complexity of the data sets as well as the data preparation techniques have a major impact on the Naïve Bayes algorithm performance in certain data sets. The Naïve Bayes method can work effectively on smaller, non-redundant data sets such the Wisconsin data collection. The Naïve Bayes method will perform less on bigger and more complicated data sets, such the METABRIC and SEER data sets. Feature selection and dimensionality reduction methods can help to lower data dimensions and sparsity, therefore enhancing the performance on challenging data sets. Concurrently, various augmentation methods include Laplace smoothing can help to reduce data sparsity issues. And loosen the independence assumption using ensemble techniques (like Bayesian networks) to enhance classification results.

d. Random Forest

The Random Forest model performs well in all data sets. It can effectively capture the characteristics and patterns of the data by building a decision tree and combining its results, avoiding the overfitting problem that a single decision tree is prone to. It can also reduce the data sparsity problem caused by one-hot encoding to a certain extent, thereby maintaining high accuracy, precision and other values. In order to further improve the performance of the random forest algorithm on complex data sets, you can consider optimizing hyperparameters (such as the number of trees, maximum depth, etc.), reducing data dimensions, and using balancing techniques (such as SMOTE) to process unbalanced data. In addition, integrating more tree models and using cross-validation methods can also further improve the classification effect and the reliability of the results.

e. XGBOOST

The model can effectively improve the performance of the model by optimizing the loss function, using regularization terms, building deeper trees, etc., adapting to complex data set characteristics, and achieving excellent classification results. Although the algorithm performs well on multiple data sets, it still has some shortcomings, especially when dealing with specific types of data sets, which may show the following problems: increased risk of overfitting, difficulty in dealing with high-dimensional sparse data, sensitivity to outliers, difficulty in parameter tuning, and unsuitability for unstructured data.

f. Gradient Boosting

Using regularization terms, improving loss functions, and creating deeper trees, gradient boosting—a method of ensemble learning—improves model performance. This approach performs well in many kinds of classification problems and can adjust to intricate data set characteristics. When handling some particular kinds of data sets, the gradient lifting method still has certain flaws, nevertheless. The Wisconsin dataset shows somewhat low feature dimension and rather homogeneous data distribution. As so, it is appropriate for the gradient lifting technique. The Coimbra dataset features more

dimensions and sparse data than the Wisconsin one. The Gradient Boosting model yields not especially good results. The models in the METABRIC and SEER datasets must be especially created to produce appropriate prediction results.

g. XGBoost+Naïve Bayes+Voting Classifier

XGBoost and Naïve Bayes are utilized as fundamental classifiers in the voting technique to ascertain the ultimate classification results. As a potent ensemble learning method, XGBoost can successfully depict the intricate data interactions. Under small data sets, the Naïve Bayes algorithm also performs rather well. The voting technique can maximize the benefits of the two methods to raise the generalization capacity and generalizing accuracy overall. After one-hot encoding, the dimension of the METABRIC and SEER datasets rises noticeably; XGBoost and Naïve Bayes could thus be susceptible to some restrictions when processing high-dimensional sparse data.

h. RGX

In this stacked ensemble model, Random Forest and Gradient Boosting models are used as base models, and XGBoost is used as a meta-model. This is a powerful machine learning method that improves the overall prediction performance by inputting the prediction results of multiple base models as features into a meta-model for training. In the four data sets, the model performed very well, surpassing all other models in three data sets and obtaining the highest value.

4.2.2 Datasets

In the study, some basic machine learning algorithms were adopted, and some integrated algorithms and hybrid algorithms were also used.

a. Wisconsin

In the study of the Wisconsin dataset, a variety of simple machine learning algorithms were used, such as Decision Tree, SVM, etc., as well as some integrated algorithms, such as Random Forest, XGBoost, etc. After data preprocessing, the data set was