

**MACHINE LEARNING APPLICATIONS IN PREDICTIVE
MODELING AND MISSING DATA IMPUTATION:
INSIGHTS FROM BREAST CANCER AND MALARIA
STUDIES**

BY
OUSMAN KHAN

A Thesis Presented to the
DEANSHIP OF GRADUATE STUDIES

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN, SAUDI ARABIA

In Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

In

APPLIED STATISTICS

DECEMBER 2023

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS
DHAHRAN 31261, SAUDI ARABIA

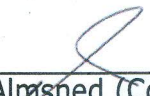
DEANSHIP OF GRADUATE STUDIES

This thesis, written by **OUSMAN KHAN** under the direction of his thesis adviser and approved by his thesis committee, has been presented to and accepted by the Dean of Graduate Studies, in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE IN APPLIED STATISTICS.**

Thesis Committee



Dr. Jimoh Olawale Ajadi (Adviser)



Dr. Fahad Almsned (Co-Adviser)



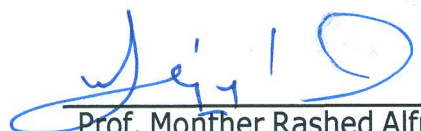
Dr. Almohanna Hani Mohammed (Member)



Dr. Emmanuel Osita Afuecheta (Member)



Dr. Mohammad Hafidz Omar (Member)



Prof. Monther Rashed Alfuraidan

Department Chairman



Prof. Suliman S. Al-Homidan

Dean of Graduate Studies

Date



To my loving family, whose unwavering support and belief in my abilities have been my greatest source of strength. Your encouragement has propelled me forward on this academic journey. I dedicate this thesis to you, with gratitude and love.

ACKNOWLEDGMENTS

I extend my deepest gratitude to my supervisor, Dr. Jimoh Ajadi, for his exceptional guidance, unwavering support, and invaluable insights throughout this research journey. His expertise and mentorship were instrumental in shaping this thesis.

I would like to take this opportunity to express my sincere gratitude to the Department of Mathematics for their continuous support and the scholarship that greatly facilitated my studies. Their unwavering assistance has been invaluable in the pursuit of my academic endeavors.

I would also like to express my sincere appreciation to the members of my thesis committee, especially, for their valuable contributions and advises in this work.

Special thanks to Dr. Hani Mohanna for generously sharing his expertise and assisting us in acquiring the breast cancer data that underpinned this study.

I am immensely grateful to Omar Ceesay for his invaluable assistance in acquiring the meteorological data. Additionally, I extend my heartfelt thanks to Fatoumatta

Jallow for her pivotal role in obtaining the essential malaria clinical data.

To my family - my parents, brothers, and sisters - I owe an immeasurable debt of gratitude. Their unwavering support, encouragement, and belief in my abilities have been my greatest source of strength.

TABLE OF CONTENTS

ACKNOWLEDGMENTS	v
LIST OF TABLES	x
LIST OF FIGURES	xii
ABSTRACT (ENGLISH)	xiv
ABSTRACT (ARABIC)	xvi
CHAPTER 1 INTRODUCTION	1
1.1 Motivation	1
1.2 Breast cancer	2
1.2.1 Breast Cancer Overview	3
1.3 Malaria	4
1.4 Missing data imputation	6
1.5 Thesis objectives and significance	9
CHAPTER 2 THEORETICAL BACKGROUND AND LITERATURE REVIEW	11
2.1 Machine Learning Techniques Employed	12
2.1.1 Decision Trees (DT)	14
2.1.2 Random Forest (RF)	18
2.1.3 K-Nearest Neighbor (KNN)	19
2.1.4 Support Vector Machines (SVM)	22

2.1.5	Artificial Neural Network (ANN)	25
2.1.6	Logistic Regression (LR)	26
2.1.7	Extreme Gradient Boosting (XGBoost)	28
2.1.8	Ensemble Modelling	29
2.2	Machine Learning Evaluation Techniques	31
2.3	Bias and Variance Trade-off in Machine Learning	35
2.4	Evaluation of Machine Learning Performance	38
2.5	Literature review	42

CHAPTER 3 PROGNOSTIC MODEL FOR PREDICTING RECURRENT IN BREAST CANCER PATIENTS IN SAUDI ARABIA 48

3.1	Objective	48
3.2	Method	49
3.2.1	Dataset used	49
3.2.2	Data Preprocessing	49
3.3	Result	54
3.4	Discussion	59

CHAPTER 4 COMPARATIVE ANALYSIS OF DIFFERENT MACHINE LEARNING TECHNIQUES IN PREDICTING BREAST CANCER RECURRENCE AMONG PATIENTS IN THE KINGDOM OF SAUDI ARABIA 68

4.1	Objective	68
4.2	Method	69
4.3	Dataset used	69
4.4	Data preprocessing	69
4.4.1	Prognostic Feature Selection	70
4.4.2	Model validation and measures of performance analysis	71
4.5	Result	73
4.6	Discussion	79

CHAPTER 5 PREDICTION OF MALARIA OUTBREAK IN EACH DISTRICT IN THE GAMBIA. 83

5.1	Objective	83
5.2	Method	83
5.2.1	Study area	83
5.2.2	Dataset used	85
5.2.3	Data Preprocessing	86
5.2.4	Performance measures	89
5.2.5	Upsampling	89
5.2.6	Model validation	91
5.3	Result	92
5.4	Discussion	97

CHAPTER 6 MACHINE LEARNING FOR PREDICTING MALARIA INCIDENCE IN EACH DISTRICT IN THE GAMBIA103

6.1	Objective	103
6.2	Method	104
6.2.1	Study area	104
6.2.2	Dataset used	104
6.2.3	Machine Learning Model: Quantile Random Forest (QRF) . . .	104
6.3	Result	105
6.3.1	Data Summary	105
6.3.2	Analysing the Influence of Climatic Variables	106
6.3.3	Prediction Model	109
6.4	Discussion	112

CHAPTER 7 INTEGRATING FAMD AND IMPUTATION METHODS IN HIGH DIMENSIONAL MIXED-DATA TYPE 115

7.1	Objective	115
7.2	Methodology	116
7.2.1	Factor Analysis for Mixed Data (FAMD)	116

7.2.2	k-Nearest Neighbors (kNN) Imputation	117
7.2.3	Random Forest Imputation	120
7.2.4	Multiple Imputation by Chained Equations (MICE)	120
7.2.5	Performance measures	121
7.3	Simulation Setup	122
7.3.1	Varying Missing Percentages in the $p - q$ features	122
7.3.2	Variations in Data Characteristics	123
7.4	Result and discussion	124
CHAPTER 8 CONCLUSION AND FUTURE DIRECTIONS		128
8.1	Conclusion	128
8.2	Future Directions	129
REFERENCES		133
VITAE		155

LIST OF TABLES

3.1	Distribution of the dependent variable	50
3.2	Medical description of the nominal variables	52
3.3	Medical description of the numerical variables in the dataset.	53
3.4	Comparison of patients' characteristics between the recurrent and non-recurrent cases.	64
3.5	Comparison of patients' characteristics between the discovery and validation cohort.	65
3.6	The result of the multivariate penalized logistic regression model shows the adjusted odd ratio with 95% CI.	66
3.7	Factors significantly associated with recurrence in multivari- able logistic regression analysis.	67
4.1	Comparison of patients' characteristics between the recurrent and non-recurrent cases in the complete dataset.	74
4.2	Performance metrics for the prediction of breast cancer recur- rence using nine machine learning techniques on the training dataset.	76
4.3	Performance metrics for the prediction of breast cancer recur- rence using nine machine learning techniques on the testing dataset.	77
4.4	Selected features in the final predicting models.	79
5.1	Summary of the variables in the dataset	89

5.2	95% confidence interval of the correlation coefficient among the climatic variables with their corresponding p-values.	93
5.3	Features included in the final predicting models.	94
5.4	Optimal parameters selected for each classifier.	95
5.5	Performance measures on the training set.	96
5.6	Performance measures on the testing set.	96
6.1	Performance Metrics for Training and Testing Datasets	109
6.2	Selected features in the QRF model.	110
7.1	Effects of different missing percentages, for p=20 covariates, $\rho = 0.9$ and percentage of categorical variables=50%.	124
7.2	Effects of different missing percentages, for p=30 covariates, $\rho = 0.9$ and percentage of categorical variables=50%.	124
7.3	Effects of different missing percentages, for p=50 covariates, $\rho = 0.9$ and percentage of categorical variables=50%.	125

LIST OF FIGURES

2.1	Machine learning chart	14
2.2	Structure of a Decision Tree	16
2.3	Assigning new data point using KNN	22
2.4	The optimal separating hyperplane in SVM.	23
2.5	Neural net with one hidden layer	27
2.6	The sigmoid shape function	27
2.7	Hold out method	33
2.8	Leave-One-Out Cross Validation in Classification Task	33
2.9	Leave-p-Out Cross Validation	34
2.10	K-Fold Cross Validation	35
2.11	Illustration of the bias and variance trade off	36
2.12	Model complexity based on prediction error	37
2.13	Diagram of a confusion matrix	41
3.1	(a) Informative features. (b) ROC curve of the logistic regression model showing the AUC values of the training and testing set.	57
3.2	Forest plot illustrating Hazard Ratios (HRs) and their corresponding confidence intervals for selected variables in the recurrence analysis.	59
4.1	Area Under the Receiver Operating Characteristic (ROC), (AUC) of all the learning algorithms.	78

5.1	The map of the Gambia showing its five main meteorological stations across different regions	84
5.2	Proportion of outbreak in both the training and testing data.	90
5.3	The Pearson’s correlation among all the climatic variables in the dataset.	92
5.4	Yearly and Monthly Variations of Malaria Cases, Deaths, and Climatic Factors.	100
5.5	Temporal Trends in Malaria Cases from 2013 to 2021.	101
5.6	This plot shows the Area Under the Receiver Operating Characteristic (ROC), (AUC) of all the learning algorithms.	102
6.1	Pearson correlation between the climatic variables and total malaria cases	107
6.2	Relationship between environmental factors and malaria cases.	108
6.3	Parity Plots for Training and Testing Sets.	111

THESIS ABSTRACT

NAME: Ousman Khan

TITLE OF STUDY: Machine Learning Applications in Predictive Modeling and Missing Data Imputation: Insights from Breast Cancer and Malaria Studies

MAJOR FIELD: Applied Statistics

DATE OF DEGREE: November, 2023

This thesis conducts a comprehensive exploration of machine learning applications, encompassing predictive modelling and innovative missing data imputation techniques. Five main chapters collectively address breast cancer recurrence prediction, malaria outbreak forecasting, malaria case prediction, and the imputation of missing values in high-dimensional mixed-data sets. In the breast cancer studies, diverse models, including multivariate logistic regression, multivariable Cox models, and various machine learning techniques, enhance recurrence prediction. Key determinants, including chemotherapy, surgery, age, and tumour size, are identified as pivotal factors. For the malaria studies, machine learning models are employed, integrating both climatic and non-climatic variables. The Random Forest model particularly stands out, demon-

strating exceptional promise for outbreak and incidence prediction. In addition, We compare three missing data imputation techniques for mixed data type by first applying Factor Analysis of Mixed Data (FAMD) for dimensionality reduction. This thesis underscores the necessity for a multidisciplinary approach to address global health challenges, merging clinical expertise with cutting-edge data-driven techniques. Given the critical nature of breast cancer and malaria, precise predictive models are imperative. The presence of missing data further underscores the importance of employing robust imputation techniques. This thesis offers valuable insights to the fields of healthcare and data analysis, providing a solid foundation for further advancements in predictive modelling and missing data imputation.

ملخص الرسالة

الاسم: أوسمان خان

عنوان الدراسة: تطبيقات التعلم الآلي في النمذجة التنبؤية ونسب البيانات المفقودة: رؤى من دراسات سرطان الثدي والمalaria

التخصص: الإحصاء التطبيقي

تاريخ الدرجة العلمية: نوفمبر، ٢٠٢٣

قوم هذه الأطروحة بإجراء استكشاف شامل لتطبيقات التعلم الآلي، بما يشمل النمذجة التنبؤية وتقنيات استكمال البيانات المفقودة المبتكرة. تتناول الفصول الرئيسية الخمسة توقعات عودة سرطان الثدي وتوقعات انتشار المalaria وتوقعات حالات المalaria واستكمال القيم المفقودة في مجموعات البيانات المختلطة ذات الأبعاد العالية. في دراسات سرطان الثدي، يعزز نماذج متنوعة، بما في ذلك الانحدار اللوجستي متعدد المتغيرات ونماذج كوكس متعددة المتغيرات وتقنيات التعلم الآلي المختلفة، التوقعات المتكررة. يتم تحديد المحددات الرئيسية، بما في ذلك العلاج الكيميائي والجراحة والعمر وحجم الورم، كعوامل حاسمة. في دراسات المalaria، يتم استخدام نماذج التعلم الآلي، مع دمج المتغيرات المناخية وغير المناخية. يبرز نموذج الغابة العشوائية بشكل خاص، مما يظهر وعودًا استثنائية لتوقعات انتشار المبيوعات والحوادث. قمنا بمقارنة ثلاث تقنيات لاستكمال البيانات المفقودة لأنواع البيانات المختلطة بتطبيق تحليل العوامل للبيانات المختلطة (FAMD) لتقليل الأبعاد. تؤكد هذه الأطروحة على ضرورة الاعتماد على نهج متعدد التخصصات لمعالجة التحديات الصحية العالمية، من خلال دمج الخبرة السريرية مع تقنيات قوية قائمة على البيانات. نظرًا للطبيعة الحرجة لسرطان الثدي والمalaria، فإن النماذج التوقعية الدقيقة ضرورية. إن وجود البيانات المفقودة يؤكد أيضًا أهمية استخدام تقنيات استكمال البيانات القوية. تقدم هذه الأطروحة نظرات قيمة في مجالات الرعاية الصحية وتحليل البيانات، مما يوفر أساسًا قويًا للتقدم في تطوير النمذجة التوقعية واستكمال البيانات المفقودة.

1 Introduction

1.1 Motivation

In the realm of global health, addressing complex challenges necessitates a multidisciplinary approach, often combining clinical expertise, advanced statistical methods, and cutting-edge technologies. Breast cancer, recognized as the most prevalent form of cancer among women worldwide, underscores the urgency for innovative approaches in its management and prognosis [1]. Concurrently, malaria continues to impose a significant health burden, particularly in regions susceptible to its transmission, mainly sub-saharan Africa [2]. Additionally, understanding and combating these horrifying diseases demand not only medical understanding but also the adept application of data-driven techniques, such as machine learning, to navigate the complexity of prediction and missing data imputation. These two challenges are further compounded by the presence of missing data, a prevalent issue in real-world datasets that can significantly impact analyses and results interpretation.

1.2 Breast cancer

Breast cancer ranks as the most frequently diagnosed cancer worldwide [3]. In 2022, it accounts for 1 in 8 cancer diagnoses and 2.3 million new instances overall in both sexes [3], [4]. In Saudi Arabia, the prevalence of breast cancer has raised alarms. As of 2005, breast cancer has constituted 19.8% of all female cancer cases in the Kingdom [5]. This trend persisted, with the Saudi Cancer Registry (SCR) identifying breast cancer as the leading cancer among Saudi women in 2016. This observation aligns with the findings of Sung et al. [3], who reported a similar trend in 2020. Breast cancer prevalence, exacerbated by numerous risk factors including age and genetic predispositions [6], underscores the imperative for precise recurrence prediction. With the increasing number of new diagnoses and survivors each year, understanding their progression, including the risk of recurrence, is crucial for improving patient outcomes [7]. Although medical advancements have led to an increased rate of successful primary treatments, a significant number of patients still face the harsh reality of recurrence [8]. This recurrent disease often has a poorer prognosis, making it more challenging to manage than the primary tumor, leading to an increased mortality rate [9]. Consequently, the ability to accurately predict breast cancer recurrence can have profound implications for patient management, prognosis, and overall survival. The traditional models, although valuable, are often constrained by their reliance on a limited set of clinical or pathological features, potentially overlooking critical individual health nuances [10]. Recognizing these limitations, this thesis embarks on a comprehensive exploration, aiming to refine predictive models for breast cancer recurrence. By employing and

comparing different machine learning techniques, we endeavor to enhance the precision of recurrence forecasts, thus paving the way for personalized patient care. Through this pursuit, we envisage a significant stride in advancing breast cancer management and patient outcomes.

1.2.1 Breast Cancer Overview

Breast cancer, which originates from the cellular components of the breast, exhibits distinct subtypes determined by its cellular origin: lobules, ducts, and connective tissue. Characterized by uncontrolled cell growth, it primarily affects women but can also occur in men. While most breast lumps are benign and non-malignant, they can increase the risk of breast cancer [11]. The exact causes of most breast cancers remain unknown; however, certain risk factors, including age, genetics, and family history, increase susceptibility[6]. Breast cancer can be detected during mammographic screenings or subsequent symptomatic presentations, such as palpable lumps [12]. Upon diagnosis, the TNM staging system, based on tumor size, lymph node involvement, and metastasis presence, guides treatment decisions [12].

The TNM staging process categorizes patients into stages based on tumor extent, with Stage 0 representing pre-cancerous Ductal Carcinoma In Situ (DCIS) and Stage *IV* indicating cancer that has spread to distant organs [12], [13]. Molecular subtyping further classifies breast cancer into luminal A, luminal B, basal-like, and HER2-enriched categories, influenced by the presence of estrogen receptors (ER), progesterone receptors (PR) and human epidermal growth factor receptor 2 (HER2) genes [14], [15]. Luminal A, representing approximately 40% of cases, exhibits favor-

able prognostics, evidenced by low 10-year recurrence rate [14], [16], [17]. In contrast, Luminal B, basal-like, and HER2+ subtypes, although less prevalent, manifest with increased recurrence and metastatic propensities [14], [15], [16]. Triple-negative breast cancer (TNBC), defined by the absence of ER, PR, and HER2, constitutes 15-20% of all breast cancer diagnoses. Its aggressive nature and suboptimal outcomes differentiate it from hormone receptor-positive cancers[18], [19], [20].

Therapeutic strategies, dependent on cancer type, stage, and extent, typically involve surgery complemented by adjuvant therapies including chemotherapy, radiotherapy, or hormone-based treatments[14], [15]. Despite therapeutic advances, recurrence remains an impending threat. Recurrence types, local, regional, and distant, differ based on lymph node and surgical margin involvement [21]. Recurrence risk varies with cancer type, stage, treatment, and extent, with most recurrences occurring within the first 2 years after diagnosis [21]. Pathological assessment determines whether distant sites harbor new cancer or represent distant metastases [21]. Accurate prognosis, monitoring, and early intervention are essential to manage the possibility of recurrence and enhance patient outcomes.

1.3 Malaria

Simultaneously, we delve into the prediction of malaria outbreaks and cases in each districts in The Gambia, drawing from historical and meteorological datasets specific to the country.

The battle against malaria, a parasitic disease transmitted primarily through the

bite of infected mosquitoes, continues to be of paramount concern, particularly in tropical and subtropical regions [2] and more so for children under the age of five [22]. Among the five parasites species that cause malaria in human beings, *P. falciparum* is the deadliest and most prevalent species in Africa. The world health organisation (WHO) reported that *P. falciparum* was responsible for 99.7% of the estimated malaria cases in Africa in 2018 [23]. It is a deadly but treatable disease. Malaria symptoms can be minor, severe, or even fatal. It includes high fevers, shivering chills, and flu-like symptoms. It is typically characterized by high recurrent fevers separated by intervals when body temperature is more or less normal [24]. An estimated 627,000 people lost their lives in 2020 due to malaria and now accounts for 7.8% of childhood mortality [25].

Over the past few decades, the Gambia's malaria burden has been significantly reduced thanks to the Gambia National Malaria Control Program (GNMCP), which plans and implements vector control strategies [26]. For instance, the number of malaria cases has decreased from 346.9 per 100,000 people in 2004 to 66 per 100,000 people in 2018 [27]. However, the eastern Gambia still has a high prevalence of malaria, with an infection rate of 10 to 40% [28] and placing a financial and health burden on households [29]. Furthermore, the disease' intensely seasonal transmission has persisted [30]. Transmission of malaria in the Gambia occurs mostly during the rainy season and the two months after it [30]. These are the seasons when the anopheles mosquito's growth and development are most favourably influenced by the weather.

The prevalence of malaria has been associated with specific climatic conditions that favors the growth of the *Plasmodium* parasite and the *Anopheles* mosquito [31].

It is reported that rainfall and temperature are the two main factors that contribute to the transmission of malaria [32]. Rainfall has a big impact on mosquitoes' lives and possibly plays a big role in the disease transmission [31]. Temperature also affects how a mosquito develops and how long it survives. Additionally, [33] also reported relative humidity alongside temperature and rainfall to have influence the plasmodium parasite's growth and maturity. In fact, the risk of malaria is reported to be reduced by a humidity level of under 60% [34].

Efforts to predict malaria outbreaks and cases have witnessed a synergy between meteorological data and advanced computational techniques [35], [36], [37]. Using the power of machine learning, we delve into the dynamics of climatic conditions in The Gambia and their profound influence on malaria transmission. By applying diverse machine learning algorithms, we aim to provide early warning systems and empower stakeholders in preemptive measures against malaria outbreaks and cases in The Gambia.

1.4 Missing data imputation

The presence of missing data casts a prevalent threat on the validity and reliability of analyses across diverse domains [38], [39], [40]. Missing data might lead to bias in the estimation of parameters and this could have a great impact on the results analysis. Additionally, it could decrease the statistical power of a test [40]. Many standard statistical methods rely on complete cases without any missing observations, so dealing with missing values is a crucial step in ensuring the validity and reliability

of any analysis. The most straightforward and widely used method for dealing with missing values is to delete all cases with missing values. However, this approach results in a loss of crucial information. An alternate method involves imputation, which entails substituting the missing data with plausible estimates.

The missing data imputation process hinges on a critical factor: the configuration of missing values. This configuration plays a pivotal role in guiding the choice of an appropriate imputation method. [41] established three classifications for missing data: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). MCAR occurs when the absence of data is unrelated to any observed or unobserved values in the dataset, suggesting no systematic variation between observations with complete and incomplete data. However, assuming true randomness in missing data is rare in practice. MAR, on the other hand, implies that data missingness is consistently linked to the observed data, often arising from recording errors, making it a more plausible scenario than MCAR [41]. NMAR, the most challenging category, arises when the likelihood of data being missing is linked to the value itself and/or other unknown data [42]. In such cases, standard methods like imputation or machine learning may not be suitable, as they presume at least MCAR or MAR. It's essential to recognize that NMAR poses significant challenges, potentially leading to biased results, whereas MCAR and MAR typically reduce statistical power [43]. Understanding these categories is crucial for ensuring the reliability and validity of statistical analyses.

Nevertheless, numerous imputation techniques have been developed to address the issue of missing values. These techniques can be categorized into two branches:

single imputation and multiple imputation [41]. Single imputation involves replacing a missing value with a single estimated value. However, this approach tends to underestimate variability as it doesn't consider uncertainty. On the other hand, multiple imputation addresses this issue by accounting for both within- and between-imputation uncertainty. When dealing with missing data, the multiple imputation method appropriately incorporates the associated uncertainties [38].

A widely employed yet computationally simple technique for handling missing data involves substituting them with the mean. However, this method neglects the interrelationships among predictive variables [44]. An alternative strategy employs machine learning algorithms such as K-Nearest Neighbors (KNN), random forest, and clustering methods for the imputation of missing values. Nevertheless, when dealing with datasets characterized by a high number of dimensions, many imputation methods face challenges due to the curse of dimensionality. In such scenarios, more effective imputation outcomes can be attained by exclusively considering pertinent covariates within the dataset. As we navigate this challenge, we integrate factor analysis of mixed data (FAMD) and imputation techniques like the K-Nearest Neighbors (KNN), random forest and Multiple Imputation by Chained Equations (MICE) to not only address the issue of missing values but also mitigate the complexities posed by high-dimensional mixed-type datasets. This innovative approach harnesses dimensionality reduction through FAMD, subsequently applying KNN-based imputation with a specialized distance measure [45], Random forest and MICE imputation methods. Through this framework, we have the objective to offer a robust framework for imputing missing values in high-dimensional datasets.

In this thesis, we embark on a multifaceted journey, weaving together insights from diverse disciplines to predict breast cancer recurrence, malaria cases and outbreak prediction, and missing data imputation. By employing various machine learning techniques, we strive to contribute meaningfully to the global pursuit of improved healthcare outcomes. Through the convergence of these endeavors, we aspire not only to refine predictive models but also to usher in a new era of data-driven healthcare interventions.

1.5 Thesis objectives and significance

In this thesis, our research encompasses the prediction of breast cancer recurrence in Saudi Arabia, utilizing a comprehensive dataset obtained from King Fahd Specialist Hospital, Dammam. Simultaneously, we delve into the prediction of malaria outbreaks and cases in each district in The Gambia, drawing from historical and meteorological datasets specific to the country. In addition to these primary objectives, we extend our focus to address the intricate challenges posed by missing data. Through the adept integration of advanced machine learning techniques, we aim to establish a comprehensive framework for increased healthcare management and prognosis. Our commitment lies in ensuring the utmost precision, accuracy, and reliability in our analyses, thereby contributing to an enhanced healthcare landscape. To summarize, the main purpose of the thesis is to:

- apply FAMD on a high dimensional dataset, then impute missing values using KNN with a special distance measure proposed in [45], Random forest and

MICE.

- use the develop technique to impute missing values in continuous, categorical and mixed data types.
- develop a high accuracy prediction model for breast cancer recurrence. This will be essential in reducing breast cancer recurrence by distinguishing precisely and in timely fashion patient at most risk of breast cancer recurrence.
- Develop a robust prediction model for malaria outbreaks and cases in The Gambia using advanced machine learning techniques. This will facilitate early warning systems and proactive measures, effectively mitigating the impact of malaria on affected communities.

Chapter 2 provides an in-depth exploration of the theoretical foundation and literature review underpinning this thesis. In Chapter 3, the focus shifts towards utilizing a logistic regression model for predicting breast cancer recurrence. Chapter 4 undertakes a comparative analysis of various machine learning techniques in the context of predicting breast cancer recurrence. Moving forward, Chapter 5 involves the application of diverse machine learning methods to forecast malaria outbreaks in each district of The Gambia. In Chapter 6, the quantile random forest (QRF) technique is employed to predict malaria cases within each district of The Gambia. Chapter 7 introduces an innovative imputation method designed for handling missing values in mixed data types. Finally, Chapter 8 summarizes the findings presented in this thesis, offering insights for future research directions and potential areas of further exploration in this field.

2 Theoretical background and literature review

The theoretical background section serves as the bedrock for this thesis, offering a comprehensive exploration of key concepts essential to understanding the methodologies employed. It encompasses:

- A detailed description on various machine learning techniques, providing a thorough grasp of their principles and applications. From concepts to theory, each technique is explained, laying the groundwork for their subsequent utilization in predicting breast cancer recurrence, malaria outbreaks, and imputing missing values.
- Exploration of the critical interplay between model complexity, prediction accuracy, and the risks of overfitting or underfitting in the section on Bias and Variance in machine learning. This nuanced understanding forms the cornerstone for creating models that strike a fine balance between capturing real-world patterns and avoiding undue complexities.
- Comprehensive coverage of the evaluation of machine learning models, a piv-

otal aspect of this study. Measures such as accuracy, sensitivity, specificity, and precision are explained, providing a robust framework for assessing model performance across diverse prediction tasks.

This section thus equips the study with a robust theoretical foundation, ensuring the wise selection and application of machine learning techniques for the specific contexts at hand.

2.1 Machine Learning Techniques Employed

In the pursuit of enhancing healthcare outcomes, the integration of advanced machine learning techniques stands as a pivotal approach. These computational methodologies hold the potential to extract complex patterns and insights from complex datasets, facilitating precise predictions and imputation of missing data. This chapter delves into the specific machine learning algorithms employed in our study, highlighting their roles in predicting breast cancer recurrence in Saudi Arabia and malaria cases in The Gambia. Furthermore, we explore how these techniques are adeptly utilized to address the ever-present challenge of missing data, ensuring a comprehensive and accurate analytical framework.

Machine learning (ML) is a subfield of artificial intelligence (AI) that uses a range of statistical, probabilistic and optimization methods that permits systems to learn and upgrade from experience [46]. Particularly in the fields of biomedical and bioinformatics, ML plays a crucial role in extracting significant insights from complex datasets [47]. Its application in healthcare holds promise for uncovering hidden patterns in

data, thereby aiding medical professionals in making early diagnoses and decisions regarding biopsies [48]. ML techniques can be put into four groups: Supervised, Unsupervised, Semi-supervised and Reinforcement learning. In supervised learning, the computer is fed with training data consisting of features (independent variables or inputs) and their corresponding dependent variable (output). Models learn patterns or trends from the training data, and utilized this learned behaviour to predict the value of the target or class variable on new data points. Unsupervised learning, on the other hand, involves feeding the computer unlabelled data without known values for the target variable. The learning algorithm autonomously identifies patterns within the data, making it particularly useful for pattern detection and descriptive modeling. Semi-supervised learning lies between supervised and unsupervised learning. It utilizes a combination of labeled and unlabeled data to train models, allowing for a more nuanced understanding of complex datasets. Reinforcement learning constitutes a unique approach wherein an agent acquires the ability to make a series of decisions through engagement with its environment. This agent receives feedback in the form of rewards or penalties, contingent on its actions, facilitating the acquisition of optimal strategies tailored to particular tasks [49]. This approach is especially powerful in scenarios where explicit training data may be scarce, and learning through trial and error is feasible.

Here is a concise overview of some of the machine learning techniques utilized in this thesis.

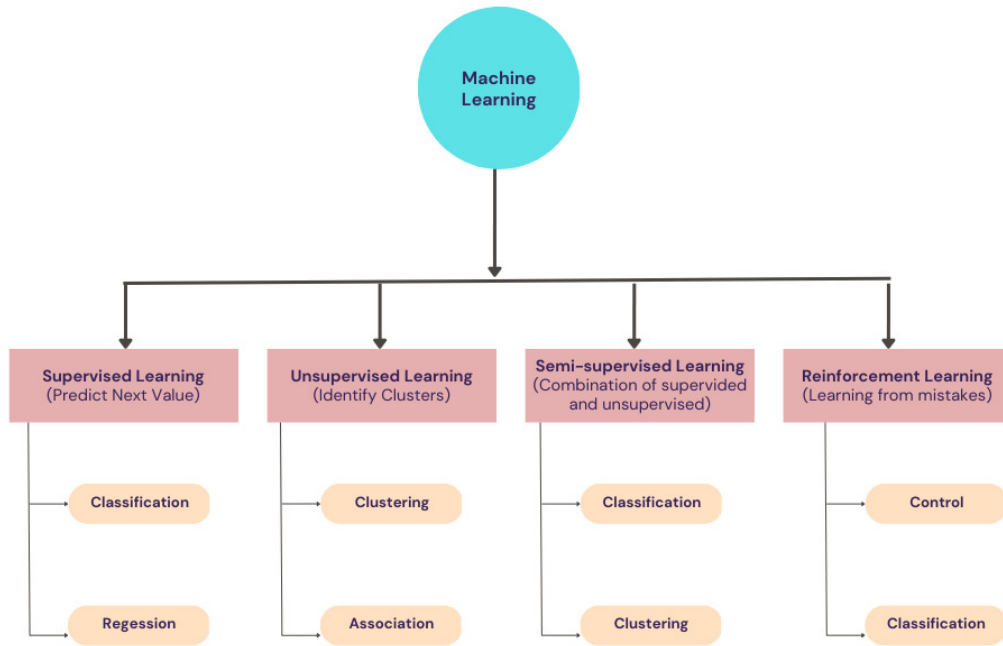


Figure 2.1: Machine learning chart

2.1.1 Decision Trees (DT)

Decision trees are widely considered in machine learning for their interpretability and strong predictive performance [50], [51]. They have found successful applications in a diverse range of classification tasks and are one of the most commonly used supervised classification techniques [52]. Notably, they can be employed for both classification and regression tasks without the need for domain-specific knowledge or parameter tuning, which adds to their simplicity and ease of understanding [52].

Structure of a Decision Tree

A decision tree (DT) is structured with leaves at the base, resulting in an inverted tree-like appearance [53]. Its components include:

- **Target or Class Variable:** This represents the outcome being predicted, such as cancer recurrence or malaria outbreak.
- **Root Nodes:** The starting point of the splitting process, determined by the variable that best divides the class variable.
- **Branches:** Connect the nodes together, forming the paths through the tree.
- **Pure/Impure Nodes:** Nodes are considered pure if they contain only one class, and impure if they have a mix of classes.
- **Decision Nodes:** Impure nodes that necessitate further splitting based on a different variable.
- **Leaf Nodes (Terminal Nodes):** These are pure nodes, signifying the end of a branch where predictions are made.

Think of a decision tree like a flowchart. It starts at the top with a question (**root node**), and depending on the answer, it follows different paths (**branches**) to more questions (**decision nodes**) until it reaches an answer (**leaf node**). Each question helps categorize the data. For instance, it might begin by asking if a tumor is malignant or benign. Based on the answer, it delves into more questions like the size of the tumor or whether it has spread to nearby lymph nodes. This process

continues until it determines a final diagnosis, like 'benign tumor with small size and no spread to lymph nodes. This employs a top-down greedy approach, as it identifies the most suitable attribute at every node from the root to the final terminal node. The objective is to determine the attribute or feature that can effectively partition the dataset at each step, commencing from the root [51].

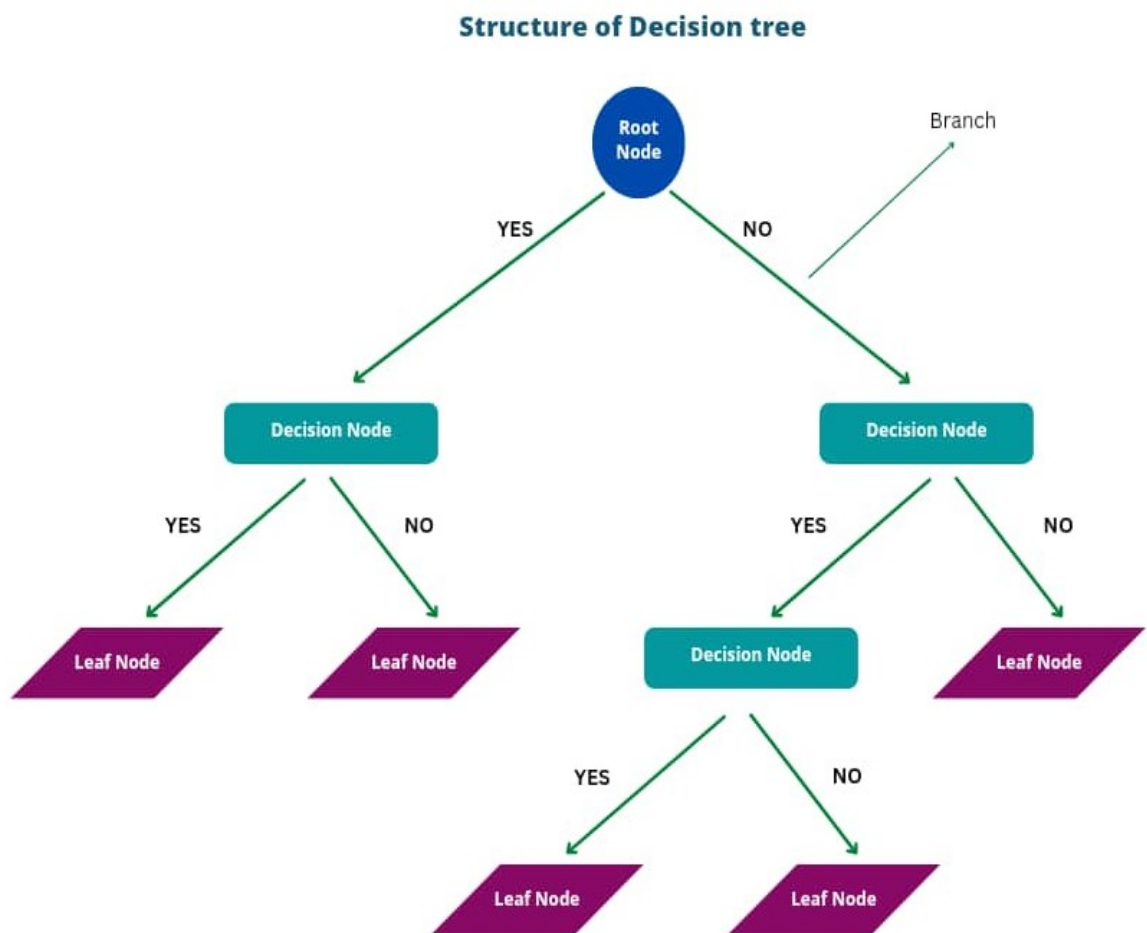


Figure 2.2: **Structure of a Decision Tree**

Attribute or feature selection

In DT, selecting the right attributes or features is crucial for effectively partitioning the data. Various metrics for attribute selection in decision trees include Entropy, Information gain, Gini index, Gain Ratio, and Reduction in Variance. In this section, we will discuss two of them: entropy and information gain.

- **Entropy:** Entropy is a concept widely used across various fields, including Decision Trees. In DT, it measures the level of impurity or randomness in data points. A node with a more mixed composition (e.g., males and females) is considered to have higher entropy than one with a more uniform distribution. Entropy ranges from 0 (pure node) to 1 (maximum entropy) [51]. The entropy of a subset S is calculated using the formula in **Eq.2.1**:

$$H(S) = - \sum_{i=1}^n p_i \log_2 p_i \quad (2.1)$$

where p_i is the probability of randomly selecting an example in class i , and S is a subset of the training examples.

- **Information Gain:** Information gain is another technique for selecting the best feature to divide the dataset. It measures the reduction in entropy achieved by partitioning the data based on a specific attribute. The attribute that yields the highest information gain is considered the best choice for splitting [51].

Information Gain is calculated using the formula in **Eq.2.2**:

$$G(S, A) = H(S) - \sum_{i=1}^n \frac{||S_v||}{||S||} H(S_v) \quad (2.2)$$

where S represents the targeted population before the split, and $||S_v||$ is the number of observations on the v_{th} split.

Decision Tree Algorithms

There are several popular decision tree algorithms, including:

- **CART (Classification and Regression Trees)**: Introduced by Breiman et al. [54], CART is widely used for both classification and regression tasks.
- **CHAID (Chi-Squared Automatic Interaction Detection)**: Proposed by Kass [55], CHAID is particularly useful in classification tasks.
- **QUEST (Quick, Unbiased, Efficient, Statistical Tree)**: Known for its efficiency, QUEST was introduced by Loh [56].
- **C4.5 and C5.0**: Developed by Quinlan [57], [58], these algorithms are renowned for their robustness and effectiveness in building decision trees.

2.1.2 Random Forest (RF)

Random Forest (RF) is a highly effective supervised machine learning technique that demonstrates exceptional efficiency in classification tasks. Unlike individual Decision Trees (DT), which can often suffer from overfitting and sensitivity to training data,

RF offers a robust solution. It employs an ensemble learning approach, combining multiple decision trees to enhance predictive accuracy [59].

In classification settings, RF determines the predicted class by aggregating the majority votes from the multiple decision trees. This collaborative decision-making process significantly improves the accuracy of predictions for new data points. Additionally, in regression tasks, RF calculates the mean of all outputs from each decision tree to arrive at the final prediction. One of the key strengths of RF lies in its ability to reduce bias and prevent overfitting [53]. It accomplishes this by introducing randomization in two critical aspects. Firstly, RF utilizes bootstrap samples of the initial data to train each tree. Secondly, it randomly selects a subset of features to determine the best split for each node. This dual randomization strategy ensures that the individual trees are uncorrelated, resulting in low variance. As a result, RF provides a high degree of stability and accuracy in its predictions, even when faced with slight variations in the training data. This feature makes it an invaluable tool in various machine learning applications.

2.1.3 K-Nearest Neighbor (KNN)

K-Nearest Neighbors (KNN) is a fundamental and easy machine learning algorithm, widely acknowledged for its simplicity and effectiveness [60], [61]. It operates on a supervised learning basis, particularly excelling in classification tasks. What sets KNN apart is its non-parametric nature, classifying test data points based on proximity to the nearest training data points, independent of the dataset's dimensions or parameters [62]. This characteristic has earned KNN the nickname of a "lazy learner" as

it preserves the training dataset rather than actively learning from it. KNN is versatile, applicable in both regression and classification tasks, though it is predominantly utilized as a classification algorithm.

The choice of the parameter 'k' in KNN is pivotal. Larger values of 'k' can lead to higher bias and lower variance, while smaller values result in higher variance but lower bias. The selection of 'k' is dependent on the input data, with datasets containing outliers or noise potentially benefiting from higher 'k' values. Cross validation serves as a valuable procedure for determining the optimal 'k' value.

Central to the conventional KNN approach is the principle of predicting the label of a test data point through a majority rule in classification tasks or averaging in regression tasks. More precisely, it predicts the label of the test data point by considering the most prevalent class among its 'k' most similar training data points in the feature space [63], [64]. To establish this similarity, KNN employs a distance metric, with the Euclidean distance shown in **Eq.2.3** being the most prevalent in the literature.

$$d(r, s) = \sqrt{\sum_{i=1}^n (s_i - r_i)^2} \quad (2.3)$$

The KNN algorithm is instrumental in various domains, finding applications in data pre-processing to impute missing values, finance, pattern recognition, and medical prognosis. It offers several advantages, including easy implementation, adaptability to accommodate new training samples, and the presence of only one parameter, 'k'. Particularly, when ample training samples are available, the nearest neighbor

algorithm can yield precise classification outcomes. However, when dealing with high-dimensional data, applying KNN can be challenging.

The steps in executing the KNN algorithm can be outlined as follows:

Selection of the number of neighbors, 'k'.

Calculation of the distances between the query point and the current data points, ordering them from smallest to largest.

Choosing the first 'k' entries from the sorted distances.

Identifying the labels of the selected 'k' entries.

Determining the mode of the 'k' labels, assigning the new data point to the category with the highest number of neighbors if the task is classification. For regression, calculating the mean or weighted average of the 'k' values and assigning the new data point a predicted value based on the average of its neighbors.

In practical terms, imagine selecting 'k' as 5, and being presented with a new data point for either classification or regression. KNN will first identify the five observations closest to the new point using a specified distance metric. These nearest neighbors form a neighborhood, which in turn influences the outcome through a majority vote for classification, or by calculating the average of target values for regression, demonstrating the algorithm's decision-making process.

As shown in **Fig.2.3**, assuming a choice of 'k' as 5, and given a new data point in black for classification, KNN will initially identify the five closest observations to the new point using a specific distance metric. This group of nearest neighbors forms

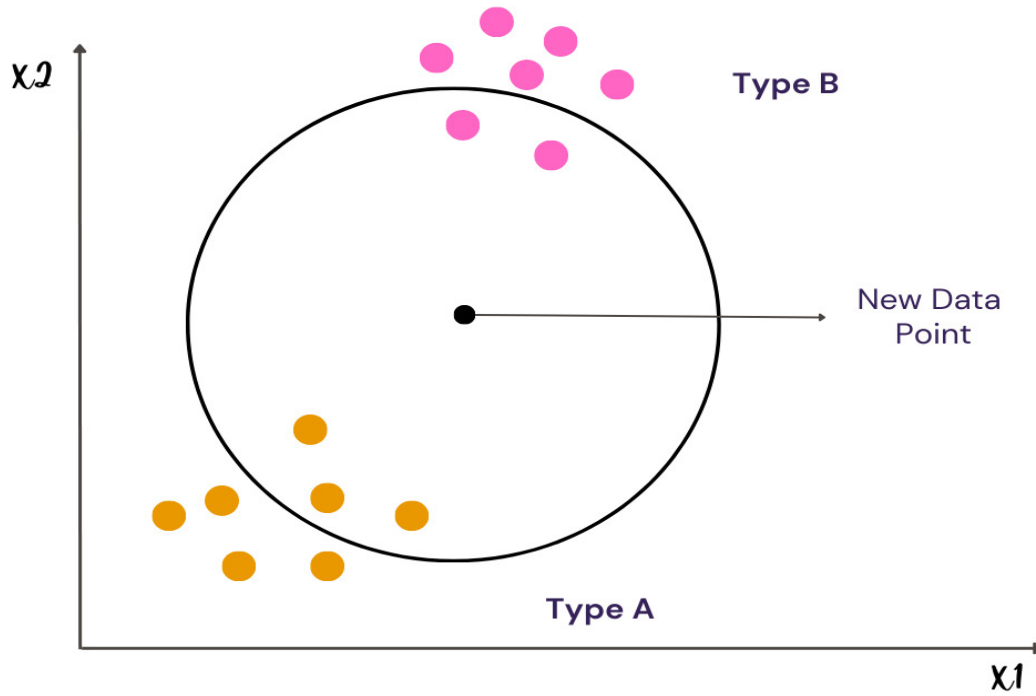


Figure 2.3: Assigning new data point using KNN

a circular neighborhood. Among these, there are 2 points of type **B** and 3 points of type **A**. Thus, adhering to the majority rule for classification, KNN will predict that the new point falls into the category of type **A**.

2.1.4 Support Vector Machines (SVM)

SVMs, introduced by Vapnik [65], are widely recognised as highly effective classifiers that demonstrates their effectiveness across various domains, including applications in the medical field [53], [66]. They are powerful tools for supervised learning tasks, encompassing both classification and regression [67]. SVMs, by design, are tailored for binary classification, and are distinguished for their exceptional predictive accuracy. They stand out as a maximum margin classification algorithm, capable of categorizing

both linear and non-linear data [68].

At the core of SVM lies the concept of a separating hyperplane, which is designed by identifying key support vectors that contribute to its definition. In an n -dimensional space, the SVM hyperplane represents the optimal boundary that separates distinct classes. The dimension of the hyperplane depends on the features within the dataset. For instance, with two features, it takes the form of a straight line; with three features, it becomes a two-dimensional plane. In binary classification, the hyperplane decisively separates the two non-overlapping outputs [53].

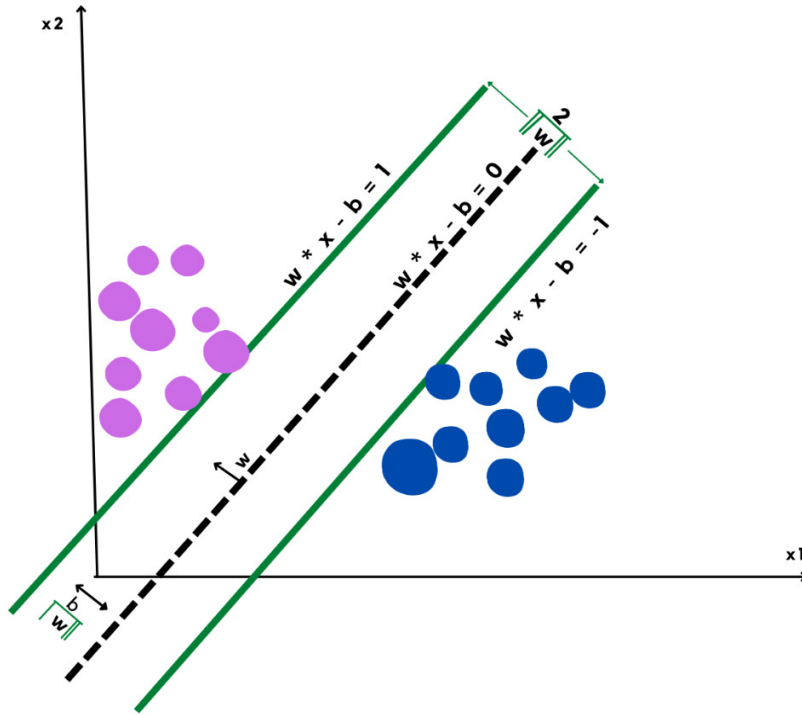


Figure 2.4: **The optimal separating hyperplane in SVM.**

The SVM algorithm operates by seeking the maximum margin hyperplane, which presents the best classification [50]. To categorize high-dimensional data, SVM offers a range of "kernels," including linear SVC, polynomial SVC, and the kernel trick. The

Kernel trick, in particular, stands out as an advanced technique for mapping data from low-dimensional to high-dimensional spaces. Selecting the appropriate SVM algorithm typically involves a trial-and-error approach, where the ideal kernel function is determined through cross-validation [53]. The penalty parameter in SVM allows for a balance between minimizing training error and maximizing the classification margin. The kernel parameters influence the complexity of the classification model and the distance between patterns in the new space. By adjusting SVM parameters, one can enhance the capability to identify the best separating hyperplane in both linear and non-linear classification scenarios, significantly impacting SVM performance [69]. An added advantage of SVM lies in its relative ease of training, its compatibility with both continuous and categorical data, and its ability to encapsulate non-linear relationships in the data. Furthermore, being a non-parametric technique, SVM operates without necessitating any assumptions about the datasets [51].

Separating Hyperplane

Let $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$ represent our training dataset of n points, where each x_i denotes a p -dimensional real vector, and the y_i are either $+1$ or -1 , corresponding to the class to which x_i belongs. The primary objective is to identify the maximum margin hyperplane that separates the x_i 's into the two distinct categories ($y_i = 1$) and ($y_i = -1$).

Any hyperplane can be expressed as the set of points satisfying:

$$w^T \cdot x - b = 0$$

Given that the training data can be separated linearly, two parallel hyperplanes can be selected to achieve this. Geometrically, the two hyperplanes are separated by a distance of $\frac{2}{||w||}$. Maximizing this distance is equivalent to minimizing $||w||$. Constraints are introduced to prevent data points from approaching the margin, resulting in the following:

$$y_i(w^T \cdot x - b) \geq 1 \quad \text{for all } 1 \leq i \leq n$$

Our objective now is to minimize $||w||$ subject to these constraints. The variables w and b dictate the classifier used to solve this problem.

In [70], the SVM technique's mathematical formulation and implementation architecture for data classification and regression are extensively covered.

2.1.5 Artificial Neural Network (ANN)

Artificial Neural Networks (ANNs) are computational models inspired by the structure of biological neural networks [71]. They consist of interconnected nodes arranged in layers: input, hidden, and output [72], [73], [74]. Every ANN has a synthetic neuron and a straightforward mathematical function as its underlying structure [75]. Each neuron in the network has the ability to take in signals, process them, and send out signals as the final product. Signals move through these layers, similar to the flow of

information in biological brains.

There are two main types of ANNs: perceptrons for basic binary classification and multilayer ANNs, such as the multi-layer perceptron (MLP) with back-propagation, for more complex tasks [71], [76]. In training, networks are set up for specific applications with initial random weights, aiming to achieve high accuracy [73]. However, understanding the internal process of ANNs can be challenging due to their "black box" nature [71]. ANNs demonstrate adaptability across various applications, from pattern recognition to data classification [72], [73], [74]. Activation functions, such as sigmoid, introduce non-linearity to the model, influencing predictions [75]. The dimensionality of the ANN depends on the dataset's features, with the weighted sum of inputs forming the basis of the activation function [75]. This mathematical model effectively captures the essence of biological neural networks [71].

The structure of neural networks can be categorized into feed-forward networks, where signals flow in one direction, and feed-back networks, which incorporate loops [72], [73], [74].

2.1.6 Logistic Regression (LR)

Logistic regression is a widely used machine learning technique for binary classification tasks, where the goal is to predict one of two class values, such as 'yes' or 'no'. It mostly employs the sigmoid function to compute probabilities between 0 and 1, effectively transforming numerical values into expressions of probability.

The sigmoid function generates an S-shaped curve, mapping any real number to a value in the range of 0 to 1. In the binary case, 0 represents the probability of an

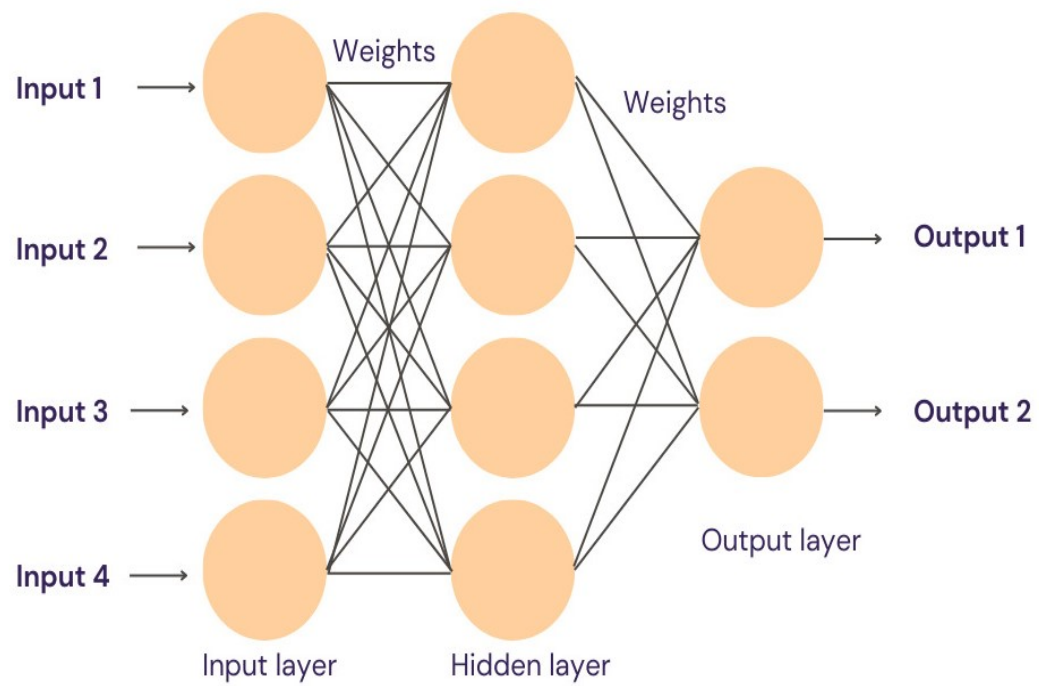


Figure 2.5: Neural net with one hidden layer

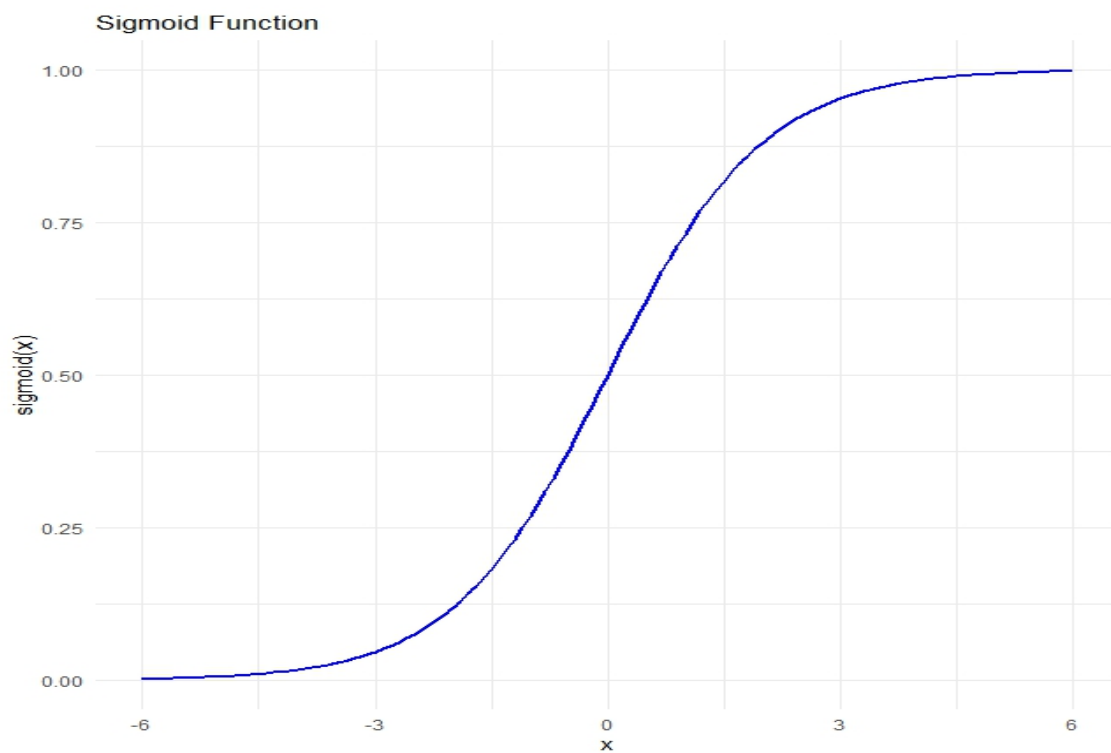


Figure 2.6: The sigmoid shape function

event not occurring, while 1 signifies the probability of it occurring. The logit, denoted as $\log\left(\frac{\theta}{1-\theta}\right)$, maps probabilities from (0,1) to $(-\infty, \infty)$, establishing a connection between independent variables and the Bernoulli distribution.

Logistic Regression is versatile and finds application in various domains, including fraud detection, disease diagnosis, and spam email identification [6], [77]. In practice, LR is employed when the dependent variable has two distinct categories, making it a powerful tool for binary classification tasks. It models the probability of occurrence rather than providing point estimates, making it especially valuable in situations where predicting likelihood is essential [6], [77].

2.1.7 Extreme Gradient Boosting (XGBoost)

XGBoost, developed by Chen and Guestrin [78], is a widely adopted supervised machine learning technique known for its high efficiency and effectiveness. It is based on ensemble trees and utilizes a scalable gradient boosting algorithm [79]. The approach of boosting is employed, where the predictions from a group of "weak" learners are combined to create a powerful "strong" learner through additive training techniques [67]. XGBoost is designed to maximize a cost objective function, which consists of a regularization term and a loss function. Additionally, it incorporates a shrinkage hyper-parameter that controls the step size of the additive expansion, helping to mitigate overfitting [80]. Notably, XGBoost demonstrates exceptional scalability, even in memory-constrained or distributed environments, making it a versatile choice for a wide range of applications.

2.1.8 Ensemble Modelling

Ensemble modelling, commonly known as ensembles, is among the best machine learning techniques. Ensemble modelling utilizes statistical methods to generate a model that yields a single forecast. More importantly, it produces a final classification or result with superior predictive accuracy by combining estimates and relying on the wisdom of the crowd. In other words, the model produced, performs better than just the base learners. The accuracy of predictions can be improved by using ensemble approaches, which are excellent for decreasing model variation. Ensemble models can be divided into various groups including sequential, parallel, homogenous, and heterogeneous.

Sequential Ensemble Techniques: Sequential ensemble techniques produce base learners sequentially. The reliance between the base learners is encouraged by the sequential generation of base learners. Adding weights to classifiers that previously misclassified data reduces prediction error for sequential ensemble models. Sequential models include AdaBoost and Gradient Boosting.

Parallel Ensemble Models: Parallel ensemble models operate simultaneously and minimize error by averaging. This method is exemplified by decision trees.

Homogeneous Ensemble: A homogeneous ensemble is one that uses multiple versions of a single technique. The base learners are of the same type and with similar qualities. A homogeneous ensemble model would consist of many decision trees cooperating to create a single forecast (**bagging**).

Heterogeneous Ensemble: Heterogeneous ensemble methods apply a variety of base learning algorithms to directly guarantee ensemble diversity. The use of k-means clustering or a neural network in conjunction with a decision tree model would be examples of a heterogeneous ensemble. These base estimators are still trained independently of each other. They are popular because they make it simple to merge existing models, enabling us to employ previously trained models as base estimators.

Ensemble modeling also has a number of subcategories; three of the most well-known are bagging, boosting, and stacking. Boosting and bagging are the most widely used ensemble methods.

Bagging: Bagging, sometimes known as "**boosted aggregating**," is an illustration of a homogeneous ensemble. It consists of two steps, **bootstrapping** and **aggregation**. The training set is subject to random sampling with replacement, a technique known as bootstrapping. This allows for the possibility of selecting the same data points multiple times. Subsequently, numerous weak models are trained separately. Depending on the specific task, such as classification or regression, the final prediction is determined by averaging or taking the majority of these individual predictions. This ensemble approach leads to a more accurate estimation. A common example of bagging is random forest. Bagging strategy is commonly used to lower predictive variance in noisy datasets.

Boosting: Boosting is a well-liked iterative technique for modifying the weight of an observation in accordance with the previous classification. It is an en-

semble method where each predictor learns from the errors of earlier predictors to improve future predictions. It tries to give an observation more weight if it was incorrectly classified. It combines a number of weak base learners that are organized sequentially. Both gradient boosting and AdaBoost are well-known instances of boosting. In general, boosting produces accurate predictive models.

Stacking: Stacking involves simultaneously running several models on the data and combining the results to create the final model. This approach is currently quite popular in machine learning competitions, including the Netflix Prize. In stacking, heterogeneous weak learners are studied, learned in parallel, and integrated through the use of a meta-learner to output a prediction based on the numerous weak learners' forecasts. This method has the drawback of having each model contribute the same amount to the ensemble forecast, regardless of how well the model did.

2.2 Machine Learning Evaluation Techniques

Machine learning models require thorough evaluation to ensure their stability and predictive performance. Cross-validation (CV) is a widely used technique for this purpose. It involves partitioning the dataset into subsets, where the model is trained on one subset and tested on another [81]. Each algorithm is trained using a portion of the data (the training sample), and the remaining portion (the validation sample) is used to calculate the algorithm's risk. The algorithm with the lowest calculated risk is then chosen through CV, indicating that the model hasn't overfitted the training

data [82]. This approach provides a valuable assurance of the model’s suitability for prediction tasks, particularly when it demonstrates robust performance on the test data, ultimately leading to accurate results [82], [83]. There are various cross-validation techniques, each with its own approach to partitioning the dataset and assessing model performance.

Hold Out Method: The Hold Out Method is a straightforward yet commonly employed evaluation technique in machine learning projects. It involves randomly dividing the dataset into two sets: a training set and a test set. Generally, more data is allocated to the training set than the test set. One of the main advantages of this method is its computational efficiency compared to other cross-validation strategies. However, it comes with drawbacks, such as high variance and bias. Changing the split can lead to varying accuracy, introducing a notable level of unpredictability [53].

Leave-One-Out Cross Validation (LOOCV): LOOCV is a classical exhaustive CV procedure closely related to the Hold Out Method. It addresses some of the limitations of the latter [53]. LOOCV involves leaving out one data point at a time for validation, utilizing the remaining data for training [82]. While LOOCV provides unbiased estimates for test-error, it often suffers from high variance due to the strong positive correlation among results from almost identical subsets of observations. Additionally, LOOCV can be computationally expensive, especially for large datasets [53].

Leave-p-Out Cross Validation (LPOCV): LPOCV is an extension of

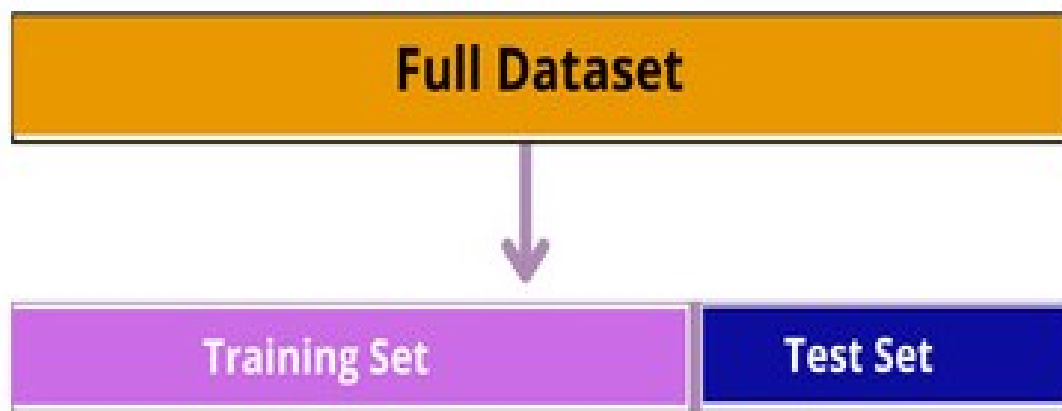


Figure 2.7: Hold out method

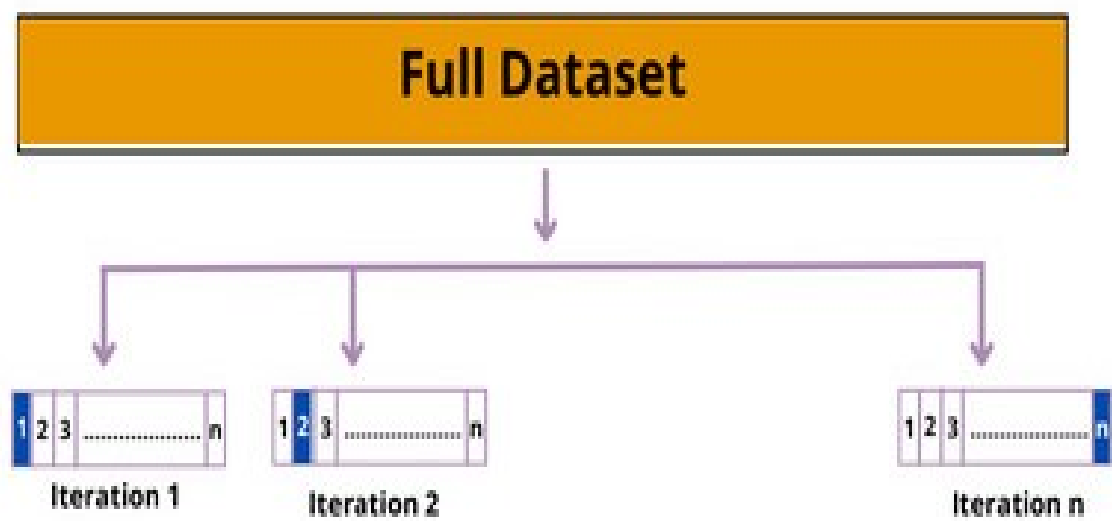


Figure 2.8: Leave-One-Out Cross Validation in Classification Task

LOOCV. It involves systematically excluding p data points from the training set and using them for validation. This method provides a balance between the computational demands of LOOCV and the Hold Out Method. However, for relatively high values of p , it may become computationally intensive [53].

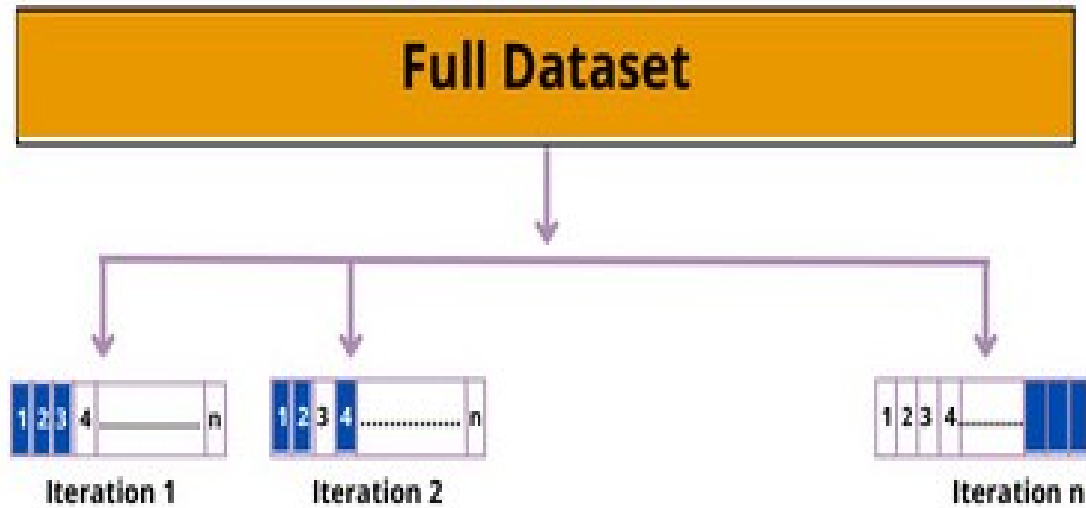


Figure 2.9: Leave-p-Out Cross Validation

K-Fold Cross Validation (K-CV): K-CV involves randomly dividing the dataset into k roughly equal-sized folds or groups. The model is trained and validated k times, with each fold serving as the validation set once. This method strikes a balance between bias and variance and often provides more accurate estimates of test error compared to LOOCV. While it requires running the model k times, making it computationally more expensive than the Hold Out Method, it is more efficient than LOOCV [53].

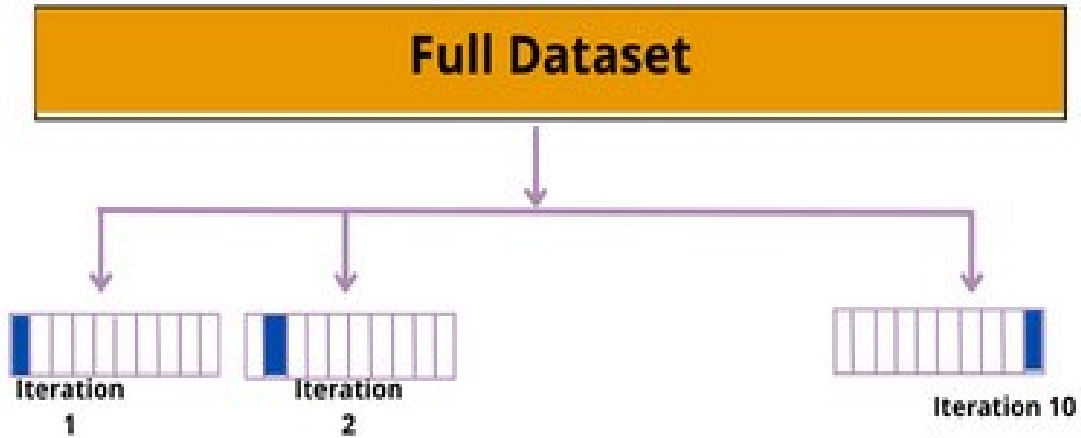


Figure 2.10: K-Fold Cross Validation

2.3 Bias and Variance Trade-off in Machine Learning

Balancing underfitting and overfitting, which determine how close a model captures the true patterns of a dataset, is a persistent challenge in machine learning. This delicate equilibrium is governed by the interplay between bias and variance. Bias signifies the disparity between an expected value and the actual value, while variance quantifies the spread of projected values.

The target shooting illustration in **Fig.2.11** strongly demonstrates the trade-off between bias and variance. The first target represents low bias, with hits closely clustered around the center, and low variance, as they are concentrated in a single area. In contrast, the second target (**labelled as B**) exhibits low bias, but high variance, with hits dispersed from one another. The third target (**labelled as C**)

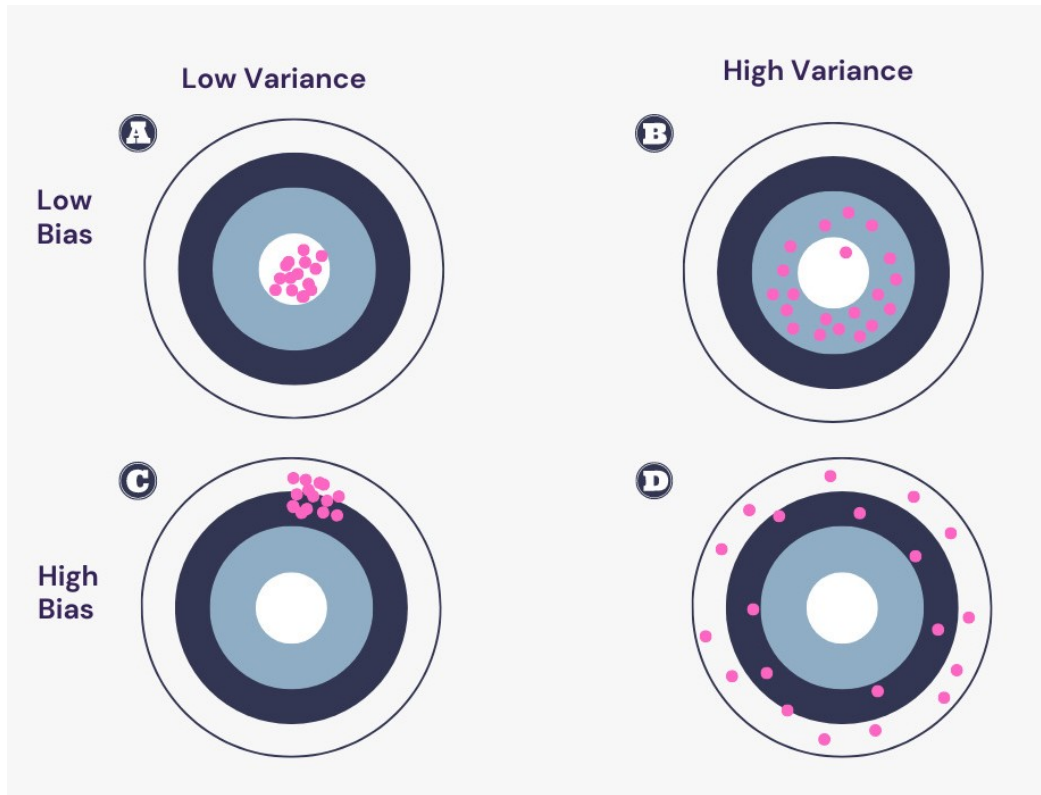


Figure 2.11: **Illustration of the bias and variance trade off**

on the second row shows high bias and low variance, while the fourth target in the same row indicates both high bias and variance. While a scenario with low bias and variance is ideal, in practice, the trade-off between optimal bias and variance is more common [84].

The graph in **Fig.2.12** demonstrates model complexity in terms of prediction error. The top line represents the training data, while the bottom line depicts the test data. Both start with a significant prediction error on the left due to low variance and strong bias. As they progress from left to right, they tend towards high variance and low bias, resulting in high prediction error for the test data and low prediction error for the training data. The optimal distribution of prediction error lies in the middle. On the left, underfitting (low variance, high bias) is evident, while on the right, overfitting

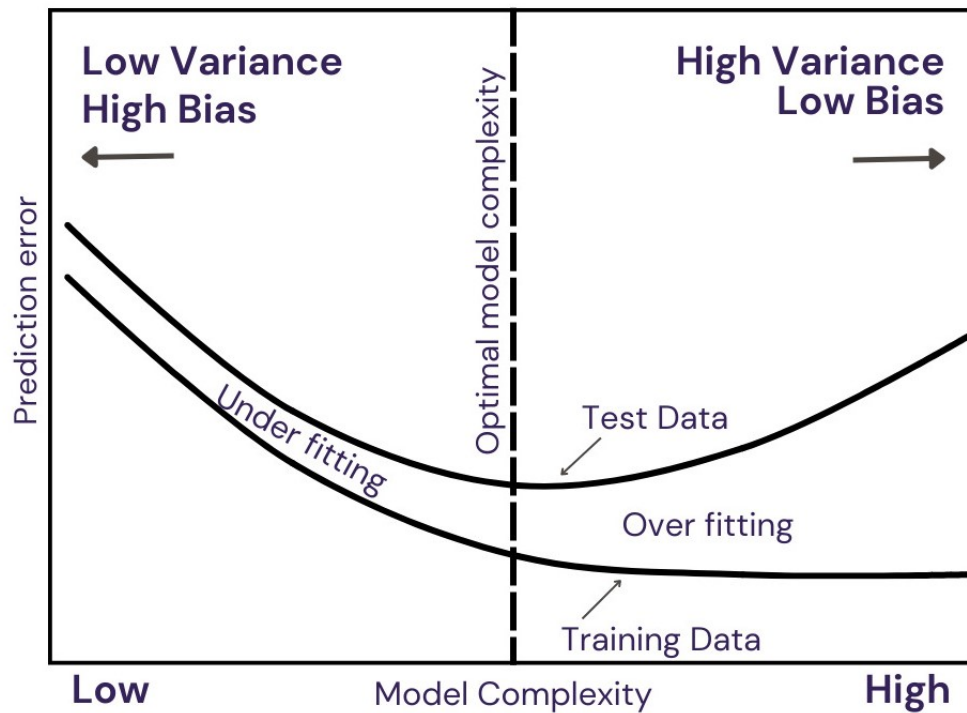


Figure 2.12: Model complexity based on prediction error

(high variance, low bias) is observed. Failure to randomize the dataset split between training and testing sets can lead to both overfitting and underfitting. Underfitting may also arise from an insufficient amount of training data to account for all potential combinations [84].

To reduce both overfitting and underfitting, adjusting the model's hyperparameters to ensure they capture patterns in both the training and testing data, and not just one set, is one approach. Regularization is another powerful method for avoiding both extremes. By penalizing an increase in a model's complexity, it artificially boosts bias error. This additional parameter serves as a warning signal, curbing excessive variation while fine-tuning the primary parameters. Cross-validation can also be employed to forestall overfitting and underfitting [84].

2.4 Evaluation of Machine Learning Performance

Evaluating a machine learning model is a crucial step in its development. It helps us understand how well the model is likely to perform on new, unseen data. One common practice in model evaluation is to split the available data into two subsets: the training set and the testing set. The training set is used to train the model, allowing it to learn the underlying patterns and relationships within the data. This is where the model adjusts its parameters based on the provided information. The testing set, on the other hand, serves as a benchmark to assess the model's performance. It contains data points that the model has never seen before. By making predictions on this unseen data and comparing them to the actual outcomes, we can gauge how well the model generalizes to new, real-world scenarios.

The choice of performance metrics depends on the nature of the task, be it classification or regression. During training, these metrics refine the classification algorithm, acting as a filter to identify optimal solutions. This process leads to more precise predictions in future assessments. In testing, the same metrics serve as evaluators, quantifying the classifier's effectiveness with unseen data [85]. For classification tasks, several performance metrics are commonly used:

Accuracy: One of the simplest and most widely used measures for machine learning is accuracy. It can be defined as the percentage of correct predictions made by a model. It is particularly popular in both binary and multi-class

classification problems.

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \quad (2.4)$$

Sensitivity: This metric, also known as recall or **true positive rate** (TPR), assesses a model's ability to correctly identify positive examples in a binary classification. A higher sensitivity implies fewer false negatives, indicating that the model effectively captures positive cases.

$$sensitivity = \frac{tp}{tp + fn} \quad (2.5)$$

Specificity (True Negative Rate): Specificity measures the proportion of true negatives that the model correctly identifies. It complements sensitivity by focusing on correctly identifying negative cases. It is also referred to as **true negative rate** (TNR).

$$specificity = \frac{tn}{tn + fp} \quad (2.6)$$

Positive Predictive Value (PPV): It is the proportion of positively classified cases which are actually positive.

$$PPV = \frac{tp}{tp + fp} \quad (2.7)$$

Negative Predictive Value (NPV): It is the proportion of negatively classified cases which are actually negative.

$$NPV = \frac{tn}{tn + fn} \quad (2.8)$$

Area under the ROC Curve (AUC-ROC): The AUC-ROC metric is employed to evaluate the effectiveness of classification models. This metric is derived from the receiver operating characteristic (ROC) curve, which illustrates the true positive rate (TPR or sensitivity) against the false positive rate (FPR or $1 - specificity$) across different classification thresholds. The AUC quantifies the entire area beneath the ROC curve in a two-dimensional space. It serves as a measure of a model's ability to differentiate between classes, with a higher AUC indicating superior performance.

These metrics are calculated based on a **confusion matrix**, where "tp," "tn," "fp," and "fn" represent true positives, true negatives, false positives, and false negatives, respectively. A Confusion matrix is a table-based illustration and one of the greatest methods for demonstrating how well machine learning programs work. The instances in each row and column of the table correspond to a predicted class and an actual class, respectively. The number of correctly recognized positive and negative examples, respectively, are shown in the confusion matrix by the letters tp and tn. The number of incorrectly identified negative and positive cases is shown by the letters fn and fp, respectively [85].

For regression tasks, different metrics are employed:



Figure 2.13: **Diagram of a confusion matrix**

Mean Absolute Error (MAE): MAE calculates the average absolute difference between predicted and actual values, providing a measure of accuracy for continuous variables.

$$MAE = \frac{\sum_{i=1}^n |y_i - \hat{y}|}{n} \quad (2.9)$$

Root Mean Square Error (RMSE): $RMSE$ is one of the most popular metrics for a regression model's machine learning performance. It is the squared average of the difference between predicted and actual values.

$$RMSE = \sqrt{\sum_{i=1}^n \frac{(y_i - \hat{y})^2}{n}} \quad (2.10)$$

The choice of performance metric depends on the specific nature of the machine learn-

ing task, whether it involves classification or regression. These metrics play a crucial role in assessing the model’s effectiveness and guiding further refinements.

2.5 Literature review

Machine learning techniques have emerged as powerful tools in the field of oncological prediction, particularly in the context of breast cancer recurrence. This area of research has accumulated substantial scholarly interest, with a notable focus on evaluating the efficacy of various machine learning techniques. While many studies have explored the prediction of breast cancer recurrence, there exists a notable gap in the utilization of datasets from Saudi Arabia.

A pivotal study by Kim et al. [86] conducted in South Korea showcased the effectiveness of Support Vector Machines (SVM) in predicting postoperative breast cancer recurrence, surpassing conventional models like St. Gallen guidelines, NPI, and Adjuvant! Online. Similarly, Boeri et al. [87] achieved commendable accuracy metrics by employing SVM and Artificial Neural Networks (ANN) for predicting postsurgical breast cancer recurrence and survival, underscoring the potential of these techniques. SVM exhibited strong predictive ability for loco-regional and systemic recurrence, achieving accuracy rates of 95.64% to 96.86%. While SVM exhibited robust predictive capacities, its sensitivity was somewhat diminished because of the class imbalances of the dataset, underscoring the challenge of skewed data distributions.

Using dataset in Iran, Ahmad et al. [88] compared SVM, Decision Trees (DT), and ANN for two-year breast cancer recurrence predictions, with SVM emerging as

superior in accuracy and sensitivity. The Surveillance, Epidemiology, and End Results (SEER) database, a popular repository, has been foundational to numerous breast cancer research. Umesh et al. [89] and Fan et al. [90] utilized the SEER datasets, with DT and C5.0 algorithms exhibiting promising outcomes.

The Wisconsin Prognosis Breast Cancer (WPBC) dataset from the UCI machine learning repository is another popular dataset among researchers. [91] compared Random Forest (RF) and Deep Neural Networks (DNN) against existing models, finding showed that RF outperformed DNN and existing models. Similarly, [92] highlighted SVM's superior prediction accuracy of SVMs compared with Naive Bayes (NB) and C4.5 DT.

[93] proposed an ensemble learning technique based on the XGBoost classifier, utilizing resampling and clustering to predict breast cancer recurrence. In a genomics-oriented study, [94] employed Autoencoders (AE) to synthesise DNA methylation data, effectively identifying genes associated with breast cancer recurrence. In a comprehensive comparison, [95] found KNN to excel in general cancer prediction, whereas SVM demonstrated superior accuracy in breast cancer recurrence prediction.

Although different machine learning methodologies have been assessed in breast cancer recurrence prediction, there is a conspicuous gap in the literature pertaining to the Saudi Arabian context. Understanding the nuances and performance of these models is imperative for personalised therapeutic strategies in breast cancer management.

Moving on to the prediction of malaria outbreaks, various studies have employed diverse machine learning approaches utilizing meteorological data and malaria cases.

Climate variables, including rainfall and temperature, have been identified as pivotal factors influencing the incidence and spread of malaria [31]. Additionally, temperature, rainfall, and relative humidity exert significant influence on the growth and maturity of the plasmodium parasite, underscoring their importance in malaria ecology [32], [96].

Kalipe et al. [33] study in Visakhapatnam, India, demonstrated the potential of Extreme Gradient Boosting (XGBoost) with 96.26% accuracy in predicting malaria outbreaks, using weather data and malaria cases collected over 6 years. Their study clearly show that weather forecast may be properly used in the future to forecast malaria epidemics and possibly take the required precautions to reduce the loss of lives from malaria. Studies integrating clinical and environmental indicators, such as Zinszer et al. [37] work in Uganda, highlighted the value of combining these data sources to enhance the accuracy of malaria forecasts.

[97], [98], [36] conducted comparative analyses of machine learning models for malaria prediction, with each study revealing valuable insights into the strengths of different approaches. These studies collectively underscored the role of non-climatic and atmospheric factors in predicting the frequency of malaria outbreaks.

The forecasting of malaria cases has been a focal point of research since the early work of Christopher's in 1911 [99]. Recent studies have utilized climatic data, including factors such as rainfall, temperature, and relative humidity, in combination with clinical data to predict malaria cases [35], [36], [37]. The integration of clinical predictors with environmental variables has been shown to enhance the accuracy of malaria forecasts in diverse endemic settings [37].

Furthermore, studies have highlighted the significance of climate variables in influencing the growth and survival of *Anopheles* mosquitoes, which play a central role in malaria transmission [31]. Factors such as temperature and rainfall have been identified as key elements contributing to the prevalence of malaria [32]. Additionally, maintaining a humidity level below 60% has been associated with reduced malaria risk [34].

Addressing the challenge of missing data is paramount in many research endeavors. Various techniques have been developed, broadly categorized into single and multiple imputation approaches [41]. Mean substitution, a computationally straightforward method, is commonly used for single value imputation. However, it may not fully capture the correlation structure among predictors [44]. Machine learning algorithms, including K-Nearest Neighbors (KNN), random forest, and clustering methods, have proven effective in imputing missing values, particularly in high-dimensional datasets [100], [101].

In high-dimensional datasets, the curse of dimensionality poses a challenge for most imputation techniques. Addressing this, [102] proposed a Bayesian lasso regression-based approach for imputing missing values in high-dimensional datasets. However, computational demands may limit its practicality as the number of variables with missing data increases. In addition, [101] introduced MissForest, an algorithm utilizing random forest for imputing missing data in high-dimensional settings. Comparative studies in [103] has demonstrated the superior efficacy of MissForest over parametric imputation methods.

The K-Nearest Neighbor (KNN) imputation method has gained popularity as a

non-parametric approach for single value imputation [100]. Studies by [100] and [104] demonstrated the effectiveness of KNNimpute, particularly in gene expression data. Different variations of the basic kNN imputation method have been proposed, all revolving around the central idea of calculating means or averages from potential neighboring data points. Tutz et al. [44] proposed a weighted nearest neighbor imputation method, offering an improvement over traditional nearest neighbor approaches, particularly in high-dimensional datasets. In a recent study, [39] also introduced a weighted nearest neighbor imputation method tailored for imputing missing data in categorical variables, demonstrating its efficacy in both simulated and real datasets. The method has been shown to perform in high dimensional dataset also.

Nevertheless, each of these techniques is constrained to exclusively handle either continuous or categorical data. There is a limited amount of literature available on the imputation of mixed data. The missForest technique, which relies on random forest and was introduced by Stekhoven and Bühlmann [101], is also well-suited for imputing missing values in mixed-type data. It has demonstrated its effectiveness in generating reliable imputations, irrespective of the number of variables, their types, and the interrelationships among them. [45] introduced a weighted version of the L_q distance specifically for mixed-type data. This method utilizes data from a selected group of crucial variables to impute missing values. Their approach, an enhancement over standard nearest neighbor imputation, exhibited superior performance, particularly in high-dimensional data scenarios. In practical applications, it's common to encounter datasets comprising a mix of continuous and nominal variables. With the surge in high-dimensional data due to increasing complexities, the need for effective

imputation methods for such mixed-type data has become increasingly crucial. Given that multiple imputation techniques falter in handling high-dimensional missing data, non-parametric single imputation methods are gaining traction [100], [45], [39], [44]. Our proposed imputation approach combines the power of Factor Analysis of Mixed Data (FAMD) with an enhanced version of the widely-used non-parametric nearest neighbors method. Specifically, we initiate the process with FAMD to reduce the dataset’s dimensionality, followed by the application of the kNN method with a customized distance function tailored for mixed data and other imputation techniques. One notable advantage of our method lies in its efficiency, as it not only saves computational resources but also eliminates irrelevant variables from consideration.

3 Prognostic Model for Predicting Recurrence in Breast Cancer Patients in Saudi Arabia

3.1 Objective

The primary objective of this chapter is to apply the logistic regression model for breast cancer recurrence within three years of treatment, utilizing data from 408 breast cancer patients at King Fahd Specialist Hospital in Dammam, Saudi Arabia. The focus is on assessing the effectiveness of logistic regression as a predictive tool for identifying patients at risk of recurrence. Recognizing the current absence of a dependable tool for predicting breast cancer recurrence in clinical settings, our study also focuses on creating an interactive application. This application will empower users to input relevant clinical variables and receive predictive probabilities, offering a dynamic resource for clinical decision-making. This tool will serve as a valuable asset for healthcare practitioners, researchers, and patients alike, aiding in the comprehensive comprehension of the identified factors.

3.2 Method

3.2.1 Dataset used

The study conducted a retrospective cohort study of female patients diagnosed with breast cancer at King Fahd Specialist Hospital in Dammam between July 2002 and November 2021. The minimum follow-up period was set at 36 months. Out of the initial dataset of 415 patient records, 7 were excluded owing to missing values, resulting in a total of 408 eligible patients for analysis. The patients were randomly partitioned into two cohorts: the discovery cohort (n=285; 70%) for the retrospective study and the validation cohort (n=123; 30%). This random allocation was implemented to mitigate potential biases and ensure that both cohorts effectively represented a broader dataset. The study protocol was rigorously reviewed and approved by the IRB of the hospital, emphasizing the research's adherence to stringent ethical guidelines, with paramount importance given to the participants' well-being and privacy.

3.2.2 Data Preprocessing

The primary outcome of interest was breast cancer recurrence within a 3-year period. It is a binary variable marked as 'Yes' for observed recurrence and 'No' for no recurrence during the follow-up period. Recurrence was assessed through a meticulous follow-up process involving diagnostic imaging, clinical examinations, and laboratory tests. The dependent variable's distribution is presented in **Table 3.1**, with values 0 indicating non-recurrence and 1 indicating recurrence cases.

To comprehensively evaluate potential predictors of recurrence, the study inte-

Class	Frequency	Proportion (%)
0 (cancer did not recur)	319	78.2%
1 (cancer did recur)	89	21.8%
total	408	100%

Table 3.1: **Distribution of the dependent variable**

grated various independent variables. These included age at diagnosis, tumor size (TS), breast segment (BS), cancer stage at diagnosis, lymph node involvement (LN), menopausal status (MS), and treatment options. Age was considered as a continuous variable, with significant implications for prognosis, tailored treatment options, and risk assessment with varying disease trajectories and responses observed across age groups. The impact of age on breast cancer outcomes has been a subject of varying perspectives among researchers. Whether or not young age at diagnosis is an adverse prognostic factor in breast carcinoma has long been controversial [105]. While it's widely believed that the disease may present unfavorable characteristics and prognosis in these young patients, some studies have reported comparable outcomes to their older counterparts [106]. It has been proposed that both young and old age may be adverse prognostic factors, but the data is inconclusive [107], [108], [109]. This variability highlights the intricate relationship between age and breast cancer prognosis. Another important continuous variable was TS, which reflects the size of the tumor in millimeters at the time of the initial diagnosis. A larger tumor size often suggests a more advanced disease stage, correlating with an increased recurrence risk [110]. The specific location of the cancer within the breast or BS also played a pivotal role in our study. Segment categorizations such as central and lower inner quadrants can shed light on the diverse manifestations of breast cancer across different breast regions

and their potential implications for recurrence. The Stage of Cancer at diagnosis is equally vital. This categorical variable, delineating disease progression levels, such as *in situ* or regional, aids in patient stratification based on the severity and spread of cancer, which is integral for treatment personalization. Our study also incorporated LN as a binary variable derived from the distant metastasis data. This indicates the presence of cancer in the lymph nodes at the time of diagnosis. Lymph node metastases are significant predictors of breast cancer prognosis [111], affecting both clinical outcomes and recurrence risk. MS was another influential factor, categorized into postmenopausal or premenopausal statuses. This has potential implications for the biological characteristics and recurrence risks associated with breast cancer.

Finally, the study delved into treatment-related factors. From the broader treatment option data, we identified specific treatments such as surgery, radiotherapy, Hormone Therapy (HT), and chemotherapy. As highlighted in previous research, including [86], these treatments have a documented influence on breast cancer recurrence rates. With this comprehensive inclusion of variables, our study enriches the existing knowledge and understanding of the factors that influence breast cancer recurrence. **Table 3.2** provides a medical context and summary of categorical predictors, while **Table 3.3** does the same for continuous predictors.

The study demonstrated thorough consideration for data quality and integrity, underscoring its commitment to conducting a robust and reliable analysis of breast cancer recurrence.

Name	Definition	Value	Proportion(%)
Lymph Node	This checks if the tumour has spread to the lymph node or not.	Yes	61.3
		No	38.7
Surgery	This determines if the patient has had surgery or not.	Yes	97.5
		No	2.5
Radiotherapy	This checks if the patient had undergone radiotherapy or not.	Radiotherapy	37.5
		No Radiotherapy	62.5
Chemotherapy	This establishes whether or not the patient has had chemotherapy.	Chemotherapy	65
		No chemotherapy	35
Hormonal therapy	This determines if the patient has had hormone therapy or not.	Hormonal therapy	30.1
		No therapy	69.9
Stage	The stage is determined by the size and extent of the cancer tumor.	In situ	4.9
		Localized	29.9
		Regional	56.1
		Distant	9.1
Menopause status	Checks the menopausal status of the patient.	Premenopausal	58.1
		Postmenopausal	41.9
Site (Breast)	Location of the cancer tumor in the breast.	LIQ	4.4
		LOQ	8.8
		UIQ	8.8
		UOQ	28.4
		Central	5.4
		Overlapping lesion	22.1
		NOS	22.1

Table 3.2: Medical description of the nominal variables in the dataset. LIQ and LUQ stands for lower inner quadrant and lower outer quadrant respectively, while UIQ and UOQ stands for upper inner quadrant and upper outer quadrant respectively. NOS stands for not otherwise specified.

Statistical Analysis

Our analysis used descriptive statistics to comprehensively summarise the dataset.

Categorical variables were presented as frequencies and percentages. Continuous vari-

Name	Definition	Mean	Range
Age at diagnosis	Age of the patient in years at the time of disease diagnosis	47.87	22-90
Tumor size	The size of the cancer tumor in <i>mm</i>	35.48	0-160

Table 3.3: **Medical description of the numerical variables in the dataset.**

ables were described using median values and their respective ranges. To compare categorical differences between the recurrence and non-recurrence groups, the chi-square test or Fisher’s exact test (when applicable) was employed. The Mann-Whitney U test was used to compare continuous variables, with the threshold for statistical significance set at $p < 0.05$.

For predictive modeling, a multivariable penalized logistic regression was applied with stratified nested cross-validations to optimize the model parameters. Sequential backward feature selection (SBFS) was used to enhance the prediction accuracy of the model for breast cancer recurrence. Of the total patients ($n= 408$), 70% (285 patients) were allocated for model training, and the remaining 30% (123 patients) were used for testing purposes. Continuous variables in the training dataset were standardized to ensure a mean of zero and a standard deviation of one. The corresponding variables in the test set were transformed using training set parameters. A critical component of our methodology was the use of multivariate Cox regression analysis. This method provides insight into multiple risk factors and their individual impact on the time until a breast cancer recurrence event.

Nested cross-validation, implemented using stratified inner and outer folds, was performed on the training dataset. The inner fold optimized the model parameters

and selected the most relevant feature subsets, while the outer fold evaluated the performance of the resultant model. To account for potential dataset imbalances, stratification was applied to both folds. Feature selection was performed using the sequential backward search selection (SBSS) approach, which began with all available features and systematically excluded the least informative features in each iteration. After determining the optimal combination of hyperparameters and feature subsets, the final model was constructed using a majority voting approach, retaining features that appeared in over 50% of all outer folds. The median of the best hyperparameters from the outer folds was then used to calibrate the final model.

The efficacy of the final model was assessed using the area under the curve (AUC) of the receiver operating characteristic curve (ROC). All computational tasks were executed in R, with the penalized logistic model being developed using the caret library (version 6.0.93). The ROC curve for the final model was plotted using the pROC library (version 1.18.0). Predictors are detailed alongside their adjusted odds ratios (AOR), confidence intervals (CI), and respective p -values.

3.3 Result

Patient Demographics

The cohort consisted of 285 female breast cancer patients, as detailed in **Table 3.4**. Specifically, 223 patients showed no recurrence, whereas 62 experienced disease recurrence. The median ages for these groups were 46 years (range:24-90 years) and 44 years (range:25-78 years), respectively, and the entire cohort had a median age

of 45 years. Notably, the age differences between the groups were not statistically significant ($p = 0.175$). When examining the Breast Segment (BS), 27.8% of the patients in the no-recurrence group had cancers in the Upper Outer Quadrant (UOQ). In contrast, the recurrence group predominantly showed cancers that were not otherwise specified (NOS), accounting for 38.7%. However, these differences in location were not statistically significant ($p = 0.062$). The median tumor sizes for patients with and without recurrence were 35 mm and 29 mm, respectively, and the median of the entire cohort was 30 mm (range:0-160 mm). The differences in tumor size between the groups were statistically significant ($p = 0.018$).

Most patients in both groups were diagnosed at the regional stage (57.0% with no recurrence and 66.1% with recurrence). The *in situ* stage was the least common in both groups. The stage differences between the groups were not statistically significant ($p = 0.253$). Regarding Lymph Node involvement (LN), 61% of the non-recurrence group had involved nodes compared to 67.7% in the recurrence group ($p = 0.41$). Both groups primarily consisted of premenopausal patients: 61% for no recurrence and 62.9% for recurrence, with no significant difference ($p = 0.899$). In terms of treatment strategies, all non-recurrence patients underwent surgery compared to 90.3% of the recurrence group, a significant difference ($p < 0.001$). The majority of the patients in both groups did not receive radiotherapy ($p = 0.343$). Most patients in either group did not undergo HT ($p = 0.78$). 72.2% of the non-recurrence group and 35.5% of the recurrence group underwent chemotherapy, a significant difference ($p < 0.001$). Surgery was the predominant treatment (229 patients), whereas HT was the least common (81 patients).

When comparing the baseline characteristics between the discovery and validation cohorts, the cohorts were largely similar (**Table 3.5**). These similarities underscore the comparability of the two cohorts.

Multivariate Penalised Logistic Regression Analysis

The model, with an area under the curve (AUC) of 76% (95% CI: 72 – 81%) and 76% (95% CI: 66 – 87%) for the training and test sets, respectively, included age, tumor size, surgery, and chemotherapy status as informative features (**Figure 6.1**). Results from the multivariate logistic regression model (**Table 3.6**) showed that receiving chemotherapy was significantly associated with the response variable, with an adjusted odds ratio (AOR) of 0.143 (95% CI: 0.089-0.218, $p < 0.0001$). This indicates that the odds of the event occurring were approximately 86% lower in the group that received chemotherapy compared to the group that did not. The surgery variable was also highly significant, with an AOR of 0.009 (95% CI: 0.005-0.014, $p < 0.0001$). This suggests that the patients who underwent surgery were approximately 99% less likely to have the event occur compared to those who did not have surgery. Tumor size (TS) was positively associated with the outcome, with an AOR of 1.485 (95% CI: 1.128-1.918, $p = 0.0043$). This suggests that for each unit increase in tumor size, the odds of the event occurring increased by approximately 48.5%. Age, on the other hand, was not significantly associated with the response variable. Although the AOR was 0.841 (95% CI: 0.657 – 1.061), the p-value of 0.1398 exceeded the commonly used threshold of 0.05 for statistical significance, suggesting that the effect of age on the response variable may be due to chance.

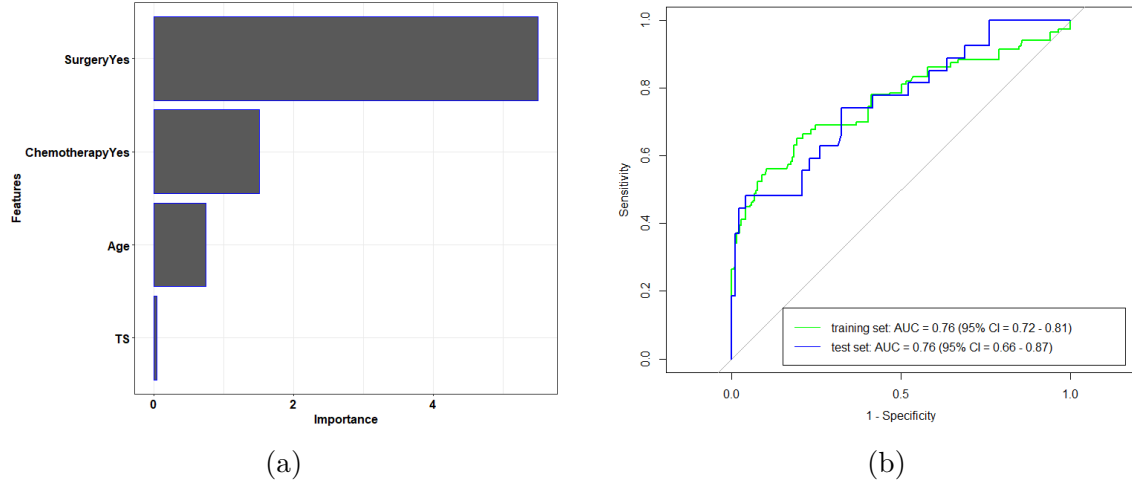


Figure 3.1: (a) Informative features. (b) ROC curve of the logistic regression model showing the AUC values of the training and testing set.

Multivariable Cox Model Analysis

This section presents the findings of our recurrence analysis using a multivariable Cox model. Initially, a model encompassing all predictors was formulated as outlined in **Table 3.7**. Age was marginally associated with breast cancer recurrence, indicating a 3% decrease in the recurrence hazard for each added year (hazard ratio [HR] = 0.97, 95% CI : 0.93 – 1.00, $p = 0.043$). Similarly, every unit growth in tumor size was correlated with a 1% increase in the recurrence hazard (HR = 1.01, 95% CI : 1.00 – 1.02, $p = 0.027$). A 95% reduction in recurrence was observed in the *in situ* cancer cases (HR = 0.05, 95% CI : 0.01 – 0.54, $p = 0.013$), and the localized cases presented a 78% reduction (HR = 0.22, 95% CI : 0.07 – 0.72, $p = 0.012$). Surgery was associated with an 80% decreased recurrence risk (HR = 0.20, 95% CI : 0.07 – 0.60, $p = 0.004$), whereas hormone therapy increased the risk by 901% (HR = 10.01, 95% CI : 2.88 – 34.73, $p < 0.001$). Chemotherapy, on the other hand, led to a substantial 95% drop in recurrence (HR = 0.05, 95% CI : 0.02 – 0.18, $p < 0.001$).

Factors such as the breast site, menopausal status, lymph node involvement, and radiotherapy were not statistically significant.

After the initial model development, we adopted a backward feature selection approach to retain informative features (**Fig.3.2**). Age significantly reduced the recurrence risk by 2.8% for each subsequent year (HR = 0.972, 95% CI : 0.949 – 0.995, $p = 0.019$). The role of tumor size was also evident, with every unit increment leading to a 1.4% increase in recurrence (HR = 1.014, 95% CI : 1.003 – 1.025, $p = 0.012$). Cancer stages revealed diverse associations. *In situ*, cases showed a 92.8% reduced recurrence likelihood (HR = 0.072, 95% CI : 0.008–0.628, $p = 0.017$), while the localized stage yielded a 70.8% reduction (HR = 0.292, 95% CI : 0.106 – 0.804, $p = 0.017$). The regional stage was not statistically significant (HR = 0.580, 95% CI : 0.229 – 1.471, $p = 0.251$). In terms of treatment, surgery contributed to an 83.4% decrease in recurrence (HR = 0.166, 95%CI : 0.060–0.461, $p = 0.001$), whereas hormone therapy amplified the risk by 982% (HR = 10.820, 95% CI : 3.129 – 37.411, $p < 0.001$). Chemotherapy showed a 95.6% decline (HR = 0.044, 95% CI : 0.013 – 0.147, $p < 0.001$). Model performance was evaluated using the C-index, yielding 0.81 for the training set (95% CI : 0.73 – 0.84) and 0.84 when tested on unseen data (95% CI : 0.76 – 0.89), highlighting the model’s efficacy and broad applicability.

Interactive Application for Breast Cancer Recurrence Risk

To enhance the accessibility and utility of our findings from the multivariate logistic regression model, we have developed an interactive application. This application allows users to input relevant clinical parameters and obtain predictive prob-

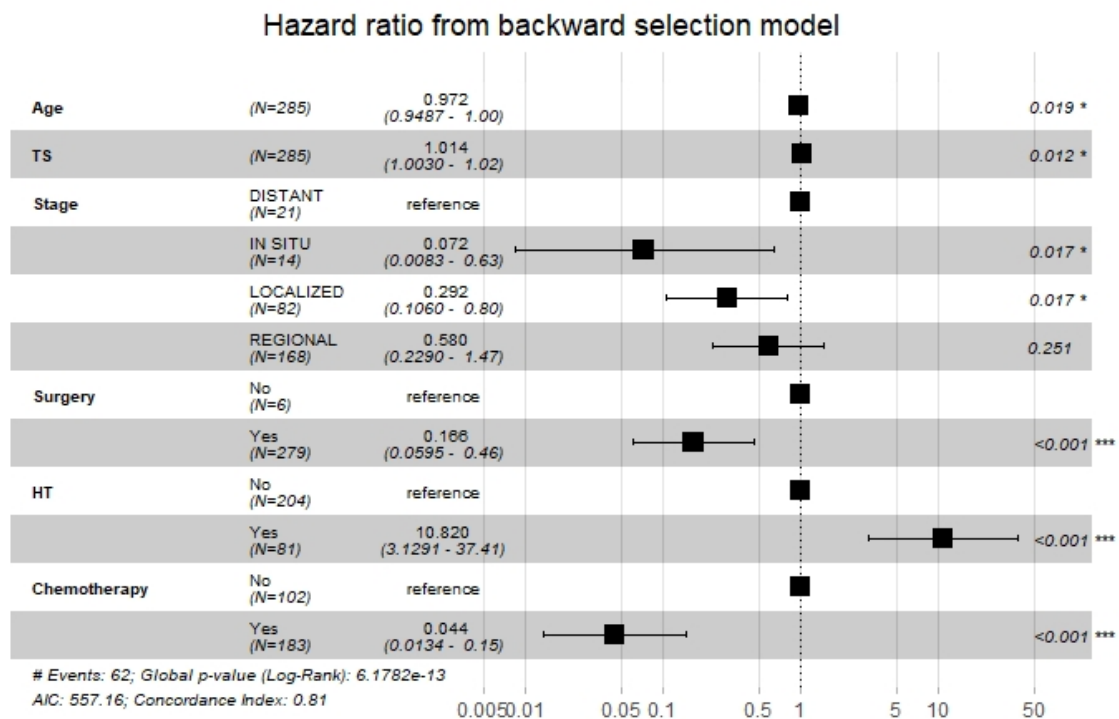


Figure 3.2: Forest plot illustrating Hazard Ratios (HRs) and their corresponding confidence intervals for selected variables in the recurrence analysis.

abilities, providing a real-time tool for clinical decision-making. Healthcare professionals, researchers, and patients can leverage this tool to better understand the factors we identified. The application is freely accessible and can be found at the following link: The application can be accessed through the link : https://iv3p9h-nurudeen-adegoke.shinyapps.io/breast_cancer

3.4 Discussion

Using clinical data from King Fahd Specialist Hospital-Dammam (KFSH-D), we developed logistic and Cox regression models to predict breast cancer recurrence. The multivariate logistic regression model demonstrated an AUC of 76% (95% CI: 72–81%) for

the training set and 76% (95% CI: 66 – 87%) for the test set. Meanwhile, Cox regression analysis achieved a C-index of 0.81 for the training set (95% CI: 0.73 – 0.84) and 0.84 for the test set (95% CI: 0.76 – 0.89). Our results showed that chemotherapy decreased the odds of recurrence by 86% (AOR:0.143, 95% CI:0.089-0.218, $p < 0.0001$), while surgery resulted in a 99% reduction in the likelihood of recurrence (AOR:0.009, 95% CI:0.005-0.014, $p < 0.0001$). The recurrence odds increased by 48.5% with an increase in tumor size (AOR:1.485, 95% CI:1.128-1.918, $p = 0.0043$), whereas age was not a significant predictor of recurrence (AOR, 0.841; 95% CI:0.657-1.061, $p = 0.1398$).

Four principal determinants: surgery, chemotherapy, age, and tumor size—were identified as pivotal predictors of breast cancer recurrence. The role of surgical interventions, explicitly breast-conserving surgery (BCS) and mastectomy, has been underscored in several studies [112], [113]. The efficacy of such procedures is contingent upon surgical precision. In cases where excision is not comprehensive, lingering malignant cells present a significant risk of recurrence. Chemotherapy is complementary in addressing potential micrometastases that are often overlooked during surgical intervention [114]. The dual strategy of surgery and chemotherapy offers a holistic approach, establishing a robust line of defense against breast cancer recurrence [114]. Age, particularly in younger patients, complicates clinical settings. Several studies have found that, while younger patients may be physiologically fit, they often grapple with aggressive cancer subtypes inherently predisposed to recurrence [115], [116]. This dictates a tailored therapeutic strategy to address unique challenges.

The salience of tumor size in our study resonates with the existing literature. A larger tumor does not merely increase cell volume but often indicates an aggressive

phenotype [117]. Such tumors suggest a prolonged phase of undetected growth or a heightened multiplication rate, which necessitates aggressive treatment paradigms [?]. Clinicians can refine therapeutic strategies and enhance patient outcomes by understanding the nuanced relationships between factors.

Our study achieved an AUC of 76% for the training set and 76% for the test set, indicating good discriminatory power in predicting breast cancer recurrence within three years of treatment. This performance aligns with several recent studies in the field. For instance, in [92], the C4.5 Decision Tree exhibited the highest AUC value of 0.745 after feature selection, outperforming Support Vector Machine (SVM) and Naive Bayes in predicting breast cancer recurrence. Nevertheless, in the context of predicting 5-year breast cancer recurrence, [118] achieved a notably higher AUC value of 0.807, surpassing our obtained result. A similar outcome was observed in the study of [119], which compared three prominent machine learning models for breast cancer survivability prognosis. The SVM, Artificial Neural Network (ANN), and Semi-Supervised Learning model (SSL) achieved AUC values of 0.80, 0.70, and 0.78, respectively. These findings suggest that our model’s predictive performance is consistent with, and in some cases superior to, previously published work in this domain.

Furthermore, it’s important to note that the performance of predictive models can be influenced by various factors including patient demographics, dataset size, and feature selection. As such, direct comparisons should be made with caution, and validation on diverse patient populations is essential for robust generalization of any predictive tool.

We designed an interactive application based on our multivariate logistic regression model results to bridge the gap between research findings and clinical applications. This user-friendly tool offers an intuitive interface where healthcare professionals, patients, and researchers can input specific clinical parameters to determine the predictive probabilities of breast cancer recurrence. It serves as a real-time decision-making tool, potentially transforming how clinicians assess recurrence risks, tailor treatments, and offer personalized advice to patients. The transparency of this digital platform, paired with its accessibility, could be a game-changer in facilitating more informed discussions between patients and healthcare providers.

Although the results of this study shed light on the potential of machine learning in predicting breast cancer recurrence, several limitations need to be acknowledged. First, the size of the dataset, consisting of only 285 observations to train the model, was relatively modest, which may have limited the broad applicability and robustness of the model. The data were exclusively collected from King Fahd Specialist Hospital-Dammam (KFSH-D) in Saudi Arabia, which may constrain the generalisability of the findings to diverse geographic and ethnic populations. Furthermore, although we analyzed ten significant features, our dataset could not include other crucial clinical, genetic, and pathological predictors. This omission may introduce bias in our predictions. Also significant is the study's strict temporal framing, defining recurrence as the disease's return within a three-year window post-treatment, potentially overlooking longer-term recurrence rates. The study also did not consider the potential co-occurrence of other cancers, which might influence breast cancer recurrence. Finally, while statistically rigorous, exclusive reliance on the logistic regression model

may overlook other models' insights or predictive capabilities.

Our study introduced a machine-learning model tailored for predicting breast cancer recurrence using a distinct dataset from a specific patient demographic. Despite its encouraging performance, the limitations emphasize the importance of further comprehensive research to validate and expand its applicability.

Table 3.4: Comparison of patients' characteristics between the recurrent and non-recurrent cases.

Predictors	No (N=223)	Yes (N=62)	Total (N=285)	P value
Age				0.175
Median	46.0	44.0	45.0	
(Range)	(24.0, 90.0)	(25.0, 78.0)	(24.0, 90.0)	
BS				0.062
CENTRAL	13 (5.8%)	1 (1.6%)	14 (4.9%)	
LIQ	12 (5.4%)	1 (1.6%)	13 (4.6%)	
LOQ	21 (9.4%)	6 (9.7%)	27 (9.5%)	
NOS	45 (20.2%)	24 (38.7%)	69 (24.2%)	
Overlapping lesion	52 (23.3%)	10 (16.1%)	62 (21.8%)	
UIQ	18 (8.1%)	3 (4.8%)	21 (7.4%)	
UOQ	62 (27.8%)	17 (27.4%)	79 (27.7%)	
TS				0.018
Median	29.0	35.0	30.0	
(Range)	(0.0, 160.0)	(9.0, 148.0)	(0.0, 160.0)	
Stage				0.253
DISTANT	15 (6.7%)	6 (9.7%)	21 (7.4%)	
IN SITU	13 (5.8%)	1 (1.6%)	14 (4.9%)	
LOCALIZED	68 (30.5%)	14 (22.6%)	82 (28.8%)	
REGIONAL	127 (57.0%)	41 (66.1%)	168 (58.9%)	
LN				0.41
No	87 (39.0%)	20 (32.3%)	107 (37.5%)	
Yes	136 (61.0%)	42 (67.7%)	178 (62.5%)	
MS				0.899
Postmenopausal	87 (39.0%)	23 (37.1%)	110 (38.6%)	
Premenopausal	136 (61.0%)	39 (62.9%)	175 (61.4%)	
Surgery				< 0.001
No	0 (0.0%)	6 (9.7%)	6 (2.1%)	
Yes	223 (100.0%)	56 (90.3%)	279 (97.9%)	
Radiotherapy				0.343
No	134 (60.1%)	42 (67.7%)	176 (61.8%)	
Yes	89 (39.9%)	20 (32.3%)	109 (38.2%)	
HT				0.78
No	161 (72.2%)	43 (69.4%)	204 (71.6%)	
Yes	62 (27.8%)	19 (30.6%)	81 (28.4%)	
Chemotherapy				< 0.001
No	62 (27.8%)	40 (64.5%)	102 (35.8%)	
Yes	161 (72.2%)	22 (35.5%)	183 (64.2%)	

LIQ and LOQ stand for lower inner and lower outer quadrants, respectively, whereas UIQ and UOQ stand for upper inner and upper outer quadrants, respectively. NOS represents not otherwise specified.

Table 3.5: Comparison of patients' characteristics between the discovery and validation cohort.

Predictors	Discovery (N=285)	Validation (N=123)	Total (N=408)	p value
response				1
NO	223 (78.2%)	96 (78.0%)	319 (78.2%)	
YES	62 (21.8%)	27 (22.0%)	89 (21.8%)	
Age				0.141
Median (Range)	45.0 (24.0, 90.0)	48.0 (22.0, 90.0)	46.0 (22.0, 90.0)	
BS				0.498
CENTRAL	14 (4.9%)	8 (6.5%)	22 (5.4%)	
LIQ	13 (4.6%)	5 (4.1%)	18 (4.4%)	
LOQ	27 (9.5%)	9 (7.3%)	36 (8.8%)	
NOS	69 (24.2%)	21 (17.1%)	90 (22.1%)	
Overlapping lesion	62 (21.8%)	28 (22.8%)	90 (22.1%)	
UIQ	21 (7.4%)	15 (12.2%)	36 (8.8%)	
UOQ	79 (27.7%)	37 (30.1%)	116 (28.4%)	
TS				0.561
Median (Range)	30.0 (0.0, 160.0)	30.0 (0.0, 160.0)	30.0 (0.0, 160.0)	
Stage				0.191
DISTANT	21 (7.4%)	16 (13.0%)	37 (9.1%)	
IN SITU	14 (4.9%)	6 (4.9%)	20 (4.9%)	
LOCALIZED	82 (28.8%)	40 (32.5%)	122 (29.9%)	
REGIONAL	168 (58.9%)	61 (49.6%)	229 (56.1%)	
LN				0.525
No	107 (37.5%)	51 (41.5%)	158 (38.7%)	
Yes	178 (62.5%)	72 (58.5%)	250 (61.3%)	
MS				0.05
Postmenopausal	110 (38.6%)	61 (49.6%)	171 (41.9%)	
Premenopausal	175 (61.4%)	62 (50.4%)	237 (58.1%)	
Surgery				0.735
No	6 (2.1%)	4 (3.3%)	10 (2.5%)	
Yes	279 (97.9%)	119 (96.7%)	398 (97.5%)	
Radiotherapy				0.717
No	176 (61.8%)	79 (64.2%)	255 (62.5%)	
Yes	109 (38.2%)	44 (35.8%)	153 (37.5%)	
HT				0.299
No	204 (71.6%)	81 (65.9%)	285 (69.9%)	
Yes	81 (28.4%)	42 (34.1%)	123 (30.1%)	
Chemotherapy				0.716
No	102 (35.8%)	41 (33.3%)	143 (35.0%)	
Yes	183 (64.2%)	82 (66.7%)	265 (65.0%)	

Table 3.6: The result of the multivariate penalized logistic regression model shows the adjusted odd ratio with 95% CI.

Predictors	AOR	95% CI	p-value
Chemotherapy			
No	1		
Yes	0.143	0.089-0.218	< 0.0001
Surgery			
No	1		
Yes	0.009	0.005 - 0.14	< 0.0001
TS (tumour size)	1.485	1.128-1.918	0.0043
Age	0.841	0.657-1.061	0.1398

Table 3.7: Factors significantly associated with recurrence in multivariable logistic regression analysis.

Variables	Hazard Ratio (95% CI)	P-value
Age	0.97 (0.93, 1.00)	0.043
BS		
CENTRAL	1.0 (reference)	
Lower inner quadrant (LIQ)	1.00 (0.06, 16.78)	0.999
Lower outer quadrant (LOQ)	1.33 (0.15, 11.58)	0.794
NOS	1.83 (0.24, 14.21)	0.564
Overlapping lesion	1.39 (0.17, 11.34)	0.756
Upper inner quadrant (UIQ)	1.34 (0.13, 13.71)	0.803
Upper outer quadrant (UOQ)	1.71 (0.22, 13.51)	0.609
TS	1.01 (1.00, 1.02)	0.027
Stage		
DISTANT	1.0 (reference)	
IN SITU	0.05 (0.01, 0.54)	0.013
LOCALIZED	0.22 (0.07, 0.72)	0.012
REGIONAL	0.78 (0.25, 2.38)	0.662
LN		
No	1.0 (reference)	
Yes	0.55 (0.16, 1.93)	0.354
MS		
Postmenopausal	1.0 (reference)	
Premenopausal	0.75 (0.36, 1.57)	0.444
Surgery		
No	1.0 (reference)	
Yes	0.20 (0.07, 0.60)	0.004
Radiotherapy		
No	1.0 (reference)	
Yes	0.77 (0.42, 1.41)	0.396
HT		
No	1.0 (reference)	
Yes	10.01 (2.88, 34.73)	< 0.0001
Chemotherapy		
No	1.0 (reference)	
Yes	0.05 (0.02, 0.18)	< 0.0001

4 Comparative Analysis of Different Machine Learning Techniques in Predicting Breast Cancer Recurrence Among Patients in the Kingdom of Saudi Arabia

4.1 Objective

This paper aims to conduct a comprehensive comparative analysis of various machine learning techniques for the prediction of breast cancer recurrence. By evaluating the performance of multiple models, the objective is to identify the most accurate and reliable predictive approach for enhancing patient outcomes. The best-performing

model will be selected to develop a practical predictive tool. This tool will serve as a valuable asset for clinicians, providing them with a reliable means to assess the risk of breast cancer recurrence in individual patients. By leveraging advanced machine learning techniques, this research endeavors to significantly improve patient care and contribute to more favorable treatment outcomes.

4.2 Method

As most studies in the literature regarding breast cancer data, this study also followed Wickham and Grolemund's [120] detailed data science life cycle methodology to analyze a set of data relevant to breast cancer patients. The most recent version of R, a well-known open-source statistical tool, was used for all of the procedures.

4.3 Dataset used

For details on the dataset, see **Dataset used** in the previous chapter.

4.4 Data preprocessing

For details on the data preprocessing, see **Data Preprocessing** in the previous chapter.

4.4.1 Prognostic Feature Selection

In the domain of predicting breast cancer recurrence, an essential aspect of methodological rigor lies in the selection of prognostic factors, ensuring that the chosen features properly capture the complex interplay of variables contributing to disease outcome. These variables can find their roots in existing knowledge or insights generously shared by medical experts. Subsequent refinement is then pursued through insightful consultations with specialists and typically followed by two closely related statistical approaches, univariate analyses based on Kaplan-Meier analysis [121] for categorical variables and univariate and multivariate Cox proportional hazard regressions for continuous variables [122]. This assesses if the feature is associated with the disease recurrence or not. However, in this case where the available dataset is limited and specific, the variable selection process can be constrained. Our dataset consisted of only 10 variables, all of which have been selected in previous studies due to their importance in predicting breast cancer recurrence. While we don't have specific references for this dataset, the significance of these variables has been highlighted in various studies within the literature. These factors include Age at Diagnosis, Tumor Size, Site (breast), Stage of Malignancy, Menopause Status, Lymph Node Involvement, Surgery, Radiotherapy, Hormone Therapy (HT), and Chemotherapy. Each of these variables has been recognized as a valuable predictor in different research contexts. This informed our decision to utilize the entire set of available features for our prediction models, as the limited dataset and the inherent importance of these variables in previous studies provided a solid basis for their inclusion.

4.4.2 Model validation and measures of performance analysis

10-fold cross validation

We used the holdout strategy to reduce overfitting and high variance in all the models. This will also provide us with accurate and reliable performance predictions for each model. This technique split the complete dataset into two random sets: 70% designated as the training set and 30% allocated as the validation set. The random division was implemented to reduce any potential bias within the prediction models. A fundamental challenge posed by the dependent variable, as highlighted in Table 3.1, was the skewed class distribution, often referred to as class imbalance. This variation was addressed by employing the up-sampling technique on the training dataset. This method uses random sampling with replacement to equalize the minority class (recurrence) to the size of the majority class (non-recurrence).

Further enhancing the model evaluation process, the widely utilized 10-fold cross-validation was applied on the training dataset. Through this method, the training data was partitioned into ten mutually exclusive folds of nearly equal size. The model underwent ten rounds of training and testing, with nine folds dedicated to training in each round and the remaining fold for testing. This strategic approach ensures the incorporation of every subset into both the training and testing phases, thereby culminating in ten distinct performance estimates such as model accuracy. To bolster the precision of the performance estimates, this 10-fold cross-validation process was iterated twenty times, a measure taken to enhance accuracy considering the dataset's relatively modest size.

Performance metrics

The study used the accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) to compare the performance of all nine models. We also use the area under the receiver operating characteristic (ROC) curve to gauge the performance of the models. Description of these performance metrics can be found in chapter 2, under the section: **Evaluation of Machine Learning Performance.**

Feature selection

Feature selection is a crucial step in disease prediction as it significantly enhances the performance of the final classifier’s prediction and clarity [123]. In the feature selection phase, we utilized the backward elimination technique to select the most impactful variables for each model. The backward elimination method is a vital technique in machine learning for refining predictive models, especially in disease prediction. It operates by systematically assessing the impact of each potential feature on the model’s performance. Beginning with the inclusion of all candidate features, the method iteratively removes the least influential feature and evaluates the model’s performance. This process continues until a predefined stopping criterion, such as a specific performance threshold or a set number of features, is reached. By progressively excluding less significant features, the method sharpens the model’s ability to generalize from training data to new, unseen data. This is critical in disease prediction, where overfitting—excessive reliance on training data—can lead to inaccurate predictions.

In addition to enhancing model precision, the backward elimination method offers computational advantages. It alleviates the computational burden, particularly

beneficial when dealing with datasets containing a large number of features. By prioritizing the most relevant variables, the method simplifies the learning process. This ensures that the resulting model not only delivers accurate predictions but does so efficiently [124]. This feature selection approach is a pivotal component of our study, contributing to the development of robust and effective models for predicting breast cancer recurrence.

4.5 Result

Upon completion of data cleaning and preparation, the final dataset encompassed 408 patients, among whom 89 experienced recurrence and 319 experienced non-recurrence within three years of cancer diagnosis. **Table 4.1** presents comprehensive descriptive statistics such as means and standard deviation for the continuous variables and percentages for the categorical variables. Moreover, the table comprehensively compares patient characteristics between instances of breast cancer recurrence and those without recurrence within the entire dataset.

A higher proportion of patients without recurrence having their cancer located in the Upper Outer Quadrant (UOQ), comprising 28.2% of the total non-recurrence cases. In contrast, for patients with recurrence, the majority of cancers were classified as not otherwise specified (NOS), accounting for 33.7% of the total recurrence cases.

The mean age of patients without recurrence stood at 48 years, while those with recurrence showed a slightly lower mean age of 45 years. This pattern aligns with the overall median age of the entire sample, which was 47 years. Patients with recurrence

Predictors	Recurrence (N=89)	Non-recurrence (N=319)	Total (N=408)
BS			
CENTRAL	3 (3.4%)	19 (6.0%)	22 (5.4%)
LIQ	1 (1.1%)	17 (5.3%)	18 (4.4%)
LOQ	7 (7.9%)	29 (9.1%)	36 (8.8%)
NOS	30 (33.7%)	60 (18.8%)	90 (22.1%)
Overlapping lesion	16 (18.0%)	74 (23.2%)	90 (22.1%)
UIQ	6 (6.7%)	30 (9.4%)	36 (8.8%)
UOQ	26 (29.2%)	90 (28.2%)	116 (28.4%)
Age	45.02+11.23	48.24+12.42	47.53+12.23
TS	39.99+26.07	34.22+23.95	35.48+24.51
Stage			
DISTANT	8 (9%)	29 (9.1%)	37 (9.1%)
IN SITU	1 (1.1%)	19 (6%)	20 (4.9%)
LOCALIZED	21 (23.6%)	101 (31.7%)	122 (29.9%)
REGIONAL	59 (66.3%)	170 (53.3%)	229 (56.1%)
LN			
No	30 (33.7%)	128 (40.1%)	158 (38.7%)
Yes	59 (66.3%)	191 (59.9%)	250 (61.3%)
MS			
Postmenopausal	34 (38.2%)	137 (42.9%)	171 (41.9%)
Premenopausal	55 (61.8%)	182 (57.1%)	237 (58.1%)
Surgery			
No	10 (11.2%)	0 (0%)	10 (2.5%)
Yes	79 (88.7%)	319 (100%)	398 (97.5%)
Radiotherapy			
No	61 (68.5%)	194 (60.8%)	255 (62.5%)
Yes	28 (31.5%)	125 (39.2%)	153 (37.5%)
HT			
No	59 (66.2%)	226 (70.8%)	285 (69.9%)
Yes	30 (33.7%)	93 (29.2%)	123 (30.1%)
Chemotherapy			
No	55 (61.8%)	88 (27.6%)	143 (35.0%)
Yes	34 (38.2%)	231 (72.4%)	265 (65.0%)

Table 4.1: Comparison of patients' characteristics between the recurrent and non-recurrent cases in the complete dataset. "LIQ" and "LOQ" correspond to the lower inner and lower outer quadrant, while "UIQ" and "UOQ" refer to the upper inner and upper outer quadrant, respectively. Additionally, "NOS" represents not otherwise specified.

exhibited a slightly larger mean tumor size of $39.99mm$ compared to $34.22mm$ among patients without recurrence. This difference in mean tumor size highlights a potential association between tumor size and recurrence risk. Additionally, the analysis of cancer stages revealed that both non-recurrence and recurrence groups predominantly comprised patients diagnosed at the '**REGIONAL**' stage, constituting 53.3% and 66.3% of cases, respectively.

A comparison of Lymph Node involvement (LN) showed that a higher percentage of patients with recurrence displayed lymph node involvement (66.3%) compared to those without recurrence (59.9%). This disparity may suggest the complex interplay of lymph node status in the context of recurrence risk. In terms of Menopausal Status (MS), premenopausal status was prevalent between both groups, encompassing 61.8% of recurrence patients and 57.1% of non-recurrence patients.

Surgical intervention exhibited a striking difference between the groups. While all non-recurrence patients underwent surgery, a significant majority of recurrence patients (88.7%) also opted for this procedure. Moreover, a higher percentage of non-recurrence patients received chemotherapy (72.4%) compared to recurrence patients (38.2%). Regarding HT, most patients did not receive this treatment in both groups. The same trend can be seen in Radiotherapy.

In summary, the comparison of patient characteristics between breast cancer recurrence and non-recurrence cases provides valuable insights into potential factors contributing to recurrence risk. The variation in tumor location, size, lymph node involvement, menopausal status, surgical choices, and chemotherapy utilization underscores the intricate nature of recurrence prediction in breast cancer cases.

With the final and clean dataset, our focus shifted towards building all the machine learning algorithms used in our study. This pivotal phase involved the development and implementation of a range of diverse models, each poised to contribute valuable insights into predicting breast cancer recurrence. We first split the entire dataset, using 70% (285 patients) for training and 30% (123 patients) for validation. Then, we applied the popular 10-fold cross-validation method to each model on the training set. In this method, nine folds are used for training, and one for testing in each iteration until each fold is used for testing. This produces 10 testing results, and the average of these 10 testing results is returned as the final estimate. However, due to the small size of the dataset, we repeated the 10-fold cross-validation 20 times, yielding 20 different performance estimates. Finally, the average of the 20 repeats of 10-fold cross-validation served as the final performance estimate. The results shown in **Table 4.2** and **Table 4.3** are achieved by using the training and testing datasets, respectively.

Models	Performance metrics				
	Accuracy	Sensitivity	Specificity	PPV	NPV
KNN	0.8105	0.9082	0.5513	0.8430	0.6935
C5.0 DT	0.9965	0.9955	1.0000	1.0000	0.9839
RF	0.8526	0.9058	0.6613	0.9058	0.6613
ANN	0.8491	0.9128	0.6418	0.8924	0.6935
svmLinear	0.7544	0.9091	0.4592	0.7623	0.7258
svmRadial	0.9789	1.0000	0.9118	0.9371	1.0000
bstTree	0.9368	0.9812	0.8056	0.9372	0.9355
XGBoost	0.9965	1.0000	0.9841	0.9955	1.0000

Table 4.2: **Performance metrics for the prediction of breast cancer recurrence using nine machine learning techniques on the training dataset.**

Results from the testing dataset, as shown in **Table 4.3**, indicate that the C5.0 DT outperformed all other models used in the study with an accuracy of 0.8049, followed

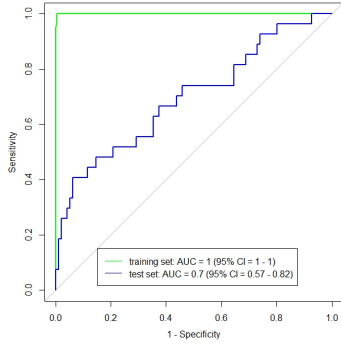
Models	Performance metrics				
	Accuracy	Sensitivity	Specificity	PPV	NPV
KNN	0.7805	0.8557	0.5000	0.8438	0.4815
C5.0 DT	0.8049	0.8273	0.6154	0.9479	0.2963
RF	0.7967	0.8447	0.5500	0.9062	0.4074
ANN	0.7886	0.8571	0.5200	0.8750	0.4815
svmLinear	0.7724	0.8778	0.4848	0.8229	0.5926
svmRadial	0.7805	0.8108	0.5000	0.9375	0.2222
bstTree	0.7886	0.8431	0.5238	0.8958	0.4074
XGBoost	0.7967	0.8447	0.5500	0.9062	0.4074

Table 4.3: **Performance metrics for the prediction of breast cancer recurrence using nine machine learning techniques on the testing dataset.**

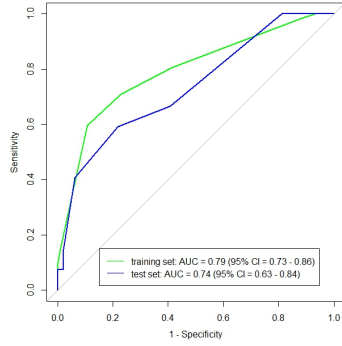
by RF and XGBoost, both achieving an accuracy of 0.7967. The C5.0 DT also obtained the highest specificity value of 0.6154. However, it had the lowest sensitivity value of 0.8273. It should be noted that both RF and XGBoost shared the same results in all metrics using the testing dataset. Although svmLinear had the lowest accuracy with a value of 0.7724, it achieved the highest sensitivity value of 0.8778, closely followed by ANN with a value of 0.8571. KNN also had a high sensitivity value of 0.8557. C5.0 DT also obtained the highest Positive Predictive Value (PPV) of 0.9479, followed by XGBoost and RF, each with 0.9062. This suggests that these models are sensitive in predicting negative tested samples. However, the Negative Predictive Values (NPV) of all the predicting models are relatively small, with 0.5926 being the highest value achieved by the svmLinear model. This suggests that all the models are less sensitive in predicting positive tested samples.

Fig.5.6 shows the ROC curve and AUC values which also compare the performance of each of the nine machine learning techniques. The ROC curve is a crucial tool for analytical test estimates.

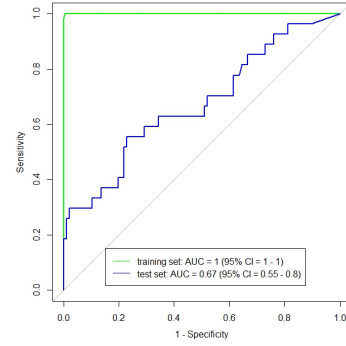
The features selected for each predicting model are presented in **Table 6.2**.



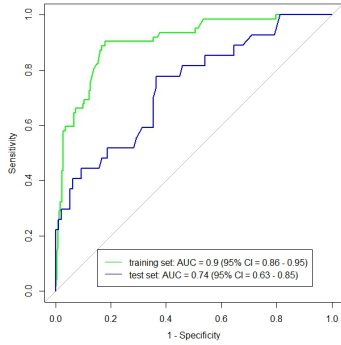
(a) XGBoost



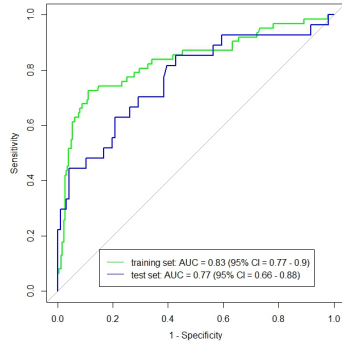
(b) KNN



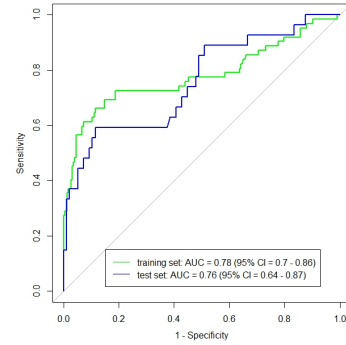
(c) C5.0 DT



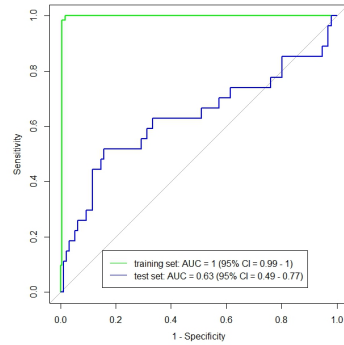
(d) RF



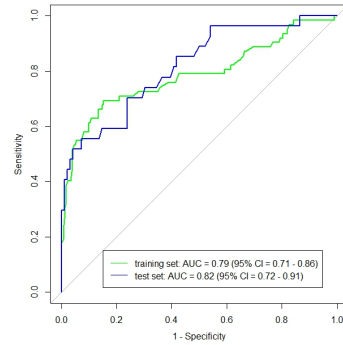
(e) ANN



(f) svmLinear



(g) svmRadial



(h) bstTree

Figure 4.1: Area Under the Receiver Operating Characteristic (ROC), (AUC) of all the learning algorithms. The green and the blue curve shows the ROC curve using the training and testing data respectively.

Features	Predictive models							
	KNN	C5.0 DT	RF	ANN	svmLinear	svmRadial	bstTree	XGBoost
Age	✓	✓	✓	✓	✓	✓	✓	✓
T-S		✓	✓	✓	✓	✓	✓	✓
L-N			✓		✓	✓		✓
Surgery	✓	✓	✓	✓	✓		✓	✓
R-T						✓		
C-T	✓	✓	✓	✓	✓	✓	✓	✓
H-T						✓		
Stage	✓	✓		✓	✓		✓	✓
M-S						✓		✓
S-B		✓				✓	✓	✓

Table 4.4: **Selected features in the final predicting models.** The inclusion of a variable in any model is shown by the tick (✓) and space left blank for any feature not selected. $T - S$, $L - N$, $M - S$ and $S - B$ represents tumor size, lymph node, menopause status and site of breast respectively. Radiotherapy, chemotherapy and hormonal therapy are represented by $R - T$, $C - T$ and $C - T$ respectively.

Chemotherapy is the only feature selected by all predicting models, while surgery is chosen by all models except svmRadial. Age at diagnosis has been included in all models except KNN. Radiotherapy is the least selected feature, appearing in only one model (svmRadial). Menopausal status also appears in only two models (svmRadial and XGBoost). XGBoost and svmRadial are the models with the highest number of selected features, each having eight.

4.6 Discussion

This study presents a comprehensive comparison of various ML algorithms with the objective of developing an accurate prognostic model for predicting breast cancer recurrence. Notably, the C5.0 DT model emerges as the most effective among the algorithms considered. Given the profound emotional impact of post-treatment relapse on patients, optimizing therapeutic strategies to prevent or delay recurrence is

of paramount importance. However, achieving this goal depends on the precise identification of individuals at increasing risk. By assessing various ML models using a dataset from Saudi Arabia, this research seeks to ascertain their predictive potential for forecasting three-year recurrence. Our results reveal that the C5.0 DT algorithm demonstrated superior performance compared to all other examined machine learning models, with XGBoost and RF following closely.

Our results align with existing literature, where SVM exhibited superior predictive performances across various studies such as [86], [87], [88], [92]. Similarly, the DT algorithm was identified as the most accurate predictor in different studies, as noted in [89], [90]. For instance, in [89], DT achieved an impressive accuracy of 94.15%, while in [90], its accuracy reached 71.17%.

The substantial performance of our best models, C5.0 DT, achieving an impressive accuracy of 0.8049 and and AUC value of 0.782, underscores the potential of machine learning algorithms in breast cancer recurrence prediction. Such high values reflect the strides made in utilizing computational methods for improved clinical decision-making.

One limitation of this study lies in its use of a relatively small dataset, potentially affecting model robustness and its ability to generalize to broader patient populations. [86] also underscore that machine learning efficacy can be compromised with limited data. To address this concern, future research endeavors could explore collaborative efforts for data sharing among multiple healthcare institutions. Additionally, leveraging transfer learning techniques with larger datasets or external validation datasets could enhance the model’s predictive performance and facilitate its application in

clinical settings.

Another limitation of this study is the potential constraint imposed by the selection of a limited set of features for prediction. While our current feature set contributes valuable insights, the model’s predictive capacity could potentially be enhanced by encompassing a more comprehensive range of relevant features. The inclusion of additional clinical, genetic, and biomarker data could provide a more nuanced understanding of the multifaceted factors influencing breast cancer recurrence. In future research, an exploration of a broader array of features could offer a more accurate representation of the underlying complexities of the disease, potentially improving the predictive accuracy of the model.

Finally, the study is also limited on relying solely on a data from a single hospital. This approach may limit the broader generalizability of our findings due to potential variations in patient demographics, treatment approaches, and data collection practices across different healthcare institutions. Future research could benefit from incorporating multi-center datasets to ensure a more comprehensive and representative assessment of the predictive model’s performance across diverse clinical settings.

In conclusion, this study report and compare a variety of machine learning techniques for predicting the recurrence of breast cancer. We developed the prediction models using a limited dataset from the KFSH-D in Saudi Arabia that consisted of 408 observations and 10 predictors after a laborious process of data cleansing and transformation. After splitting the data into training and validation set, we applied the 10-fold cross-validation technique on the training set. To address the class-imbalance

in the data, we applied the up-sampling technique. The final findings showed that the svmRadial performed best, with the maximum accuracy value of 0.8374 and an AUC value of 0.76. However, ANN has a slightly better AUC value of 0.77, which is also the best among all models. C5.0 DT came in second with an accuracy of 0.8049, followed closely by RF and XGBoost, all of which had an accuracy of 0.7967.

This study has shown that machine learning methods can accurately predict the recurrence of breast cancer. Due to its great predictive performance, the C5.0 DT model can be utilized and may be crucial in reducing breast cancer recurrence by accurately and quickly identifying patients who are most at risk. However, this model is not meant to replace the expertise of medical professionals, but to serve as a complement to their efforts in the fight against breast cancer. A good breast cancer recurrence prediction model will decrease the inconvenience and expense of patients needing to visit clinics frequently.

Additional research should be done to enhance the efficacy of these classification techniques by adding more variables and deciding on a longer follow-up period. This research could also be improved further using large dataset.

5 Prediction of malaria Outbreak in each district in The Gambia.

5.1 Objective

This chapter is dedicated to the prediction of malaria outbreaks in different districts in The Gambia. The primary objective is to utilize advanced machine learning techniques to establish a robust predictive model that facilitates early warning systems and proactive measures. This chapter aims to investigate the influence of climate variables on malaria outbreaks, providing valuable insights into the environmental factors that contribute to the prevalence of malaria in the region. Additionally, the model can be used to effectively mitigate the impact of malaria on affected communities.

5.2 Method

5.2.1 Study area

With an estimated population of 2.6 million people, The Gambia is a small and narrow nation whose borders follow the winding course of the Gambia river. The

country is surrounded by Senegal on three sides, and its westernmost point is 80 kilometers of Atlantic Ocean shoreline. The Gambia's coordinates are latitude $13^{\circ}28'N$ and longitude $16^{\circ}34'W$ respectively. The Gambia has a tropical savannah climate, which is characterized by high temperatures, sunshine, dryness, and moderate winds throughout the year. The dry season lasts from November to May, and the rainy season, which is short, typically occurs from June to October. During the dry season, temperatures range from 18 to $30^{\circ}C$, while during the wet season, temperatures range from 23 to $33^{\circ}C$. The average annual rainfall ranges from 700 millimeters in the north to 1,000 millimeters in the south and southeast. The country also has an average relative humidity of 68%.

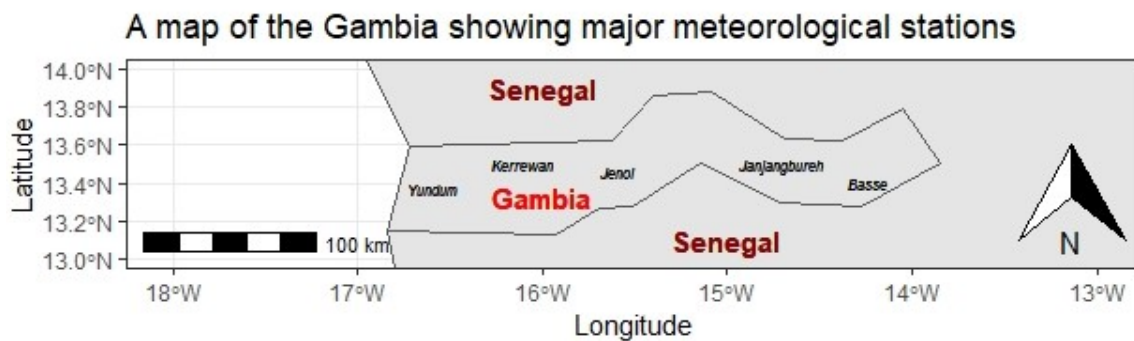


Figure 5.1: The map of the Gambia showing its five main meteorological stations across different regions

Malaria is a seasonal disease in The Gambia, with September and November seeing the highest incidence of clinical cases and fatalities [125]. The transmission of malaria in the Gambia typically commences in July or August and persists through December, with minimal to no clinical cases during the dry season [126]. The eastern region of the Gambia, particularly the Upper River Region (URR), has the highest prevalence of malaria. This area exhibits a pronounced seasonal transmission of malaria, low

vector densities, and high vector survival rates, with parous rates ranging from 81-91% compared to 27-71% in other regions [127].

5.2.2 Dataset used

This study utilized two different datasets, clinical and climatic data to predict malaria outbreak.

Clinical Data: The clinical dataset was obtained from the health management information system (HMIS) under the Directorate of Planning and Information at the Ministry of Health. The Directorate collects and reports Malaria cases from health facilities across the country, which are classified as uncomplicated or severe. Uncomplicated cases allow the patient to walk, sit, stand, be fully alert, drink well, and not be pale. Severe cases, caused by *Plasmodium falciparum* infection, can result in life-threatening consequences. Inappropriate dosing, inefficient medication use, or delayed treatment of uncomplicated malaria can lead to severe cases. The Health Management Information System (HMIS) under the Directorate of Planning and Information at the Ministry of Health compiles monthly malaria cases for each district in The Gambia from January 2013 to December 2021, including population size data for each district. It contains monthly records of malaria cases, including severe and uncomplicated cases, as well as total deaths for each district. These records were collected by qualified healthcare professionals from district hospitals and clinics across the country. The data spanned from January 2013 to December 2021.

Meteorological Data: The Department of Water Resources, under the Ministry of Fisheries, Water Resources, and National Assembly Matters, monitored and managed climatic data from January 2013 to December 2021. The collected data included total rainfall, relative humidity, and minimum and maximum temperature for each district, measured in degree Celsius, percent, and millimeters, respectively. These readings were taken from nine weather stations dispersed throughout the country, ensuring that each region had at least one weather station.

5.2.3 Data Preprocessing

After combining the clinical and meteorological datasets, the final dataset contains 4428 observations and nine variables. Rigorous steps were taken to ensure the dataset's integrity and reliability. Fortunately, there were no missing values for crucial variables, such as deaths, severe cases, uncomplicated cases, and population size. However, approximately 4% of the meteorological data, including minimum temperature, maximum temperature, rainfall, and relative humidity, had missing observations. To address these missing values, we employed the mean imputation technique, calculating the mean of non-missing values for the respective month across the years 2013 to 2021. This meticulous imputation approach aimed to maintain data quality and consistency.

Moreover, the dataset underwent a comprehensive normalization process to enhance the effectiveness of machine learning algorithms. It is a critical preprocessing step that plays a pivotal role in enhancing the quality and effectiveness of machine

learning algorithms. In this study, normalization was applied to numerical independent variables using the min-max normalization technique. This process transforms features into a shared range, mitigating the impact of larger numeric values dominating the model's learning process.

Normalization becomes particularly crucial in scenarios where the scale of numerical features varies significantly. Without normalization, features with larger numeric values may exert undue influence on the learning algorithm, potentially overshadowing the contributions of smaller-scale features. The rationale behind data normalization lies in its ability to create a level playing field for numerical features, ensuring that each contributes proportionally to the learning process. This contributes to the overall robustness and reliability of the machine learning models employed in predicting malaria outbreaks. The normalization formula employed in this study, as depicted in Eq.5.1, ensures that all numerical variables are scaled to values between 0 and 1.

$$\frac{X - X_{min}}{X_{max} - X_{min}} \quad (5.1)$$

Where X_{max} and X_{min} represent the variables' maximum and minimum values, respectively. After scaling, all the numerical variables have values between 0 and 1.

Several researchers have emphasized the importance of data normalization in improving data quality and overall algorithm performance. By implementing this pre-processing technique, biases associated with feature scales are minimized, allowing machine learning models to more effectively discern patterns and relationships within the data [128].

In addition to addressing missing values and normalizing features, the dataset was carefully examined for potential anomalies and outliers. The final step involved creating the dependent variable, named "outbreak," based on a carefully designed ratio using malaria cases and population size as shown in Eq. 5.2.

$$r = \frac{C_s + C_u}{P} \quad (5.2)$$

where, C_s and C_u represent the severe and uncomplicated cases, respectively and P denotes the population size in each district in the Gambia. To categorize the ratio into meaningful groups, an outbreak was defined as a ratio of cases greater than 0.01, signifying a significant threat in this context. Conversely, a no outbreak was defined as a ratio of cases less than or equal to 0.01.

To predict the incidence of malaria in each district (**discussed in the next chapter**), we combined the severe cases with the uncomplicated cases, and utilized the remaining variables as predictors. The equation for total cases is represented as:

$$Total\ cases = C_s + C_u \quad (5.3)$$

where C_s and C_u signify the severe and uncomplicated cases respectively.

The combination of these steps not only addressed missing data but also ensured the overall reliability and quality of the dataset. This thorough preprocessing lays a strong foundation for the subsequent machine learning analyses.

The number of total cases was not considered as an independent predictor in the forecast of malaria outbreaks. Similarly, the occurrence of outbreaks was not utilized

as an independent predictor in the prediction of malaria cases. **Table 5.1** shows the summary of the variables in the dataset.

Categorical variable name	Distinct classes		
Month	12		
District	41		
Continuous variable name	Mean	S.D	Range
Total cases	281.9968	1041.171	0 – 24320
Total rainfall	77.80595	129.1016	0 – 651.4
Minimum temperature	20.13897	3.585981	10 – 37.5
Maximum temperature	35.64957	2.814196	30.2 – 47.36667
Relative humidity	56.42773	19.0511	17 – 98
Population size	53074.01	88917.42	3789 – 528789

Table 5.1: **Summary of the variables in the dataset**

5.2.4 Performance measures

We assessed each classification model’s performance using a variety of performance metrics, including sensitivity, specificity, accuracy, and AUC. Description of these performance measures can be found in **Evaluation of Machine Learning Performance** in chapter 2.

5.2.5 Upsampling

The proportion of outbreak in both the training and testing data (after splitting the data randomly) is the same as shown in **Fig.5.2**. 87 % of both the training and testing dataset is "no outbreak" cases creating a class imbalance. This can have a negative effect on any model that is fitted. We used the up-sampling method on the training dataset to fix this problem. This method uses a random sample with replacement to make the minority class (outbreak) the same size as the majority class (no outbreak).

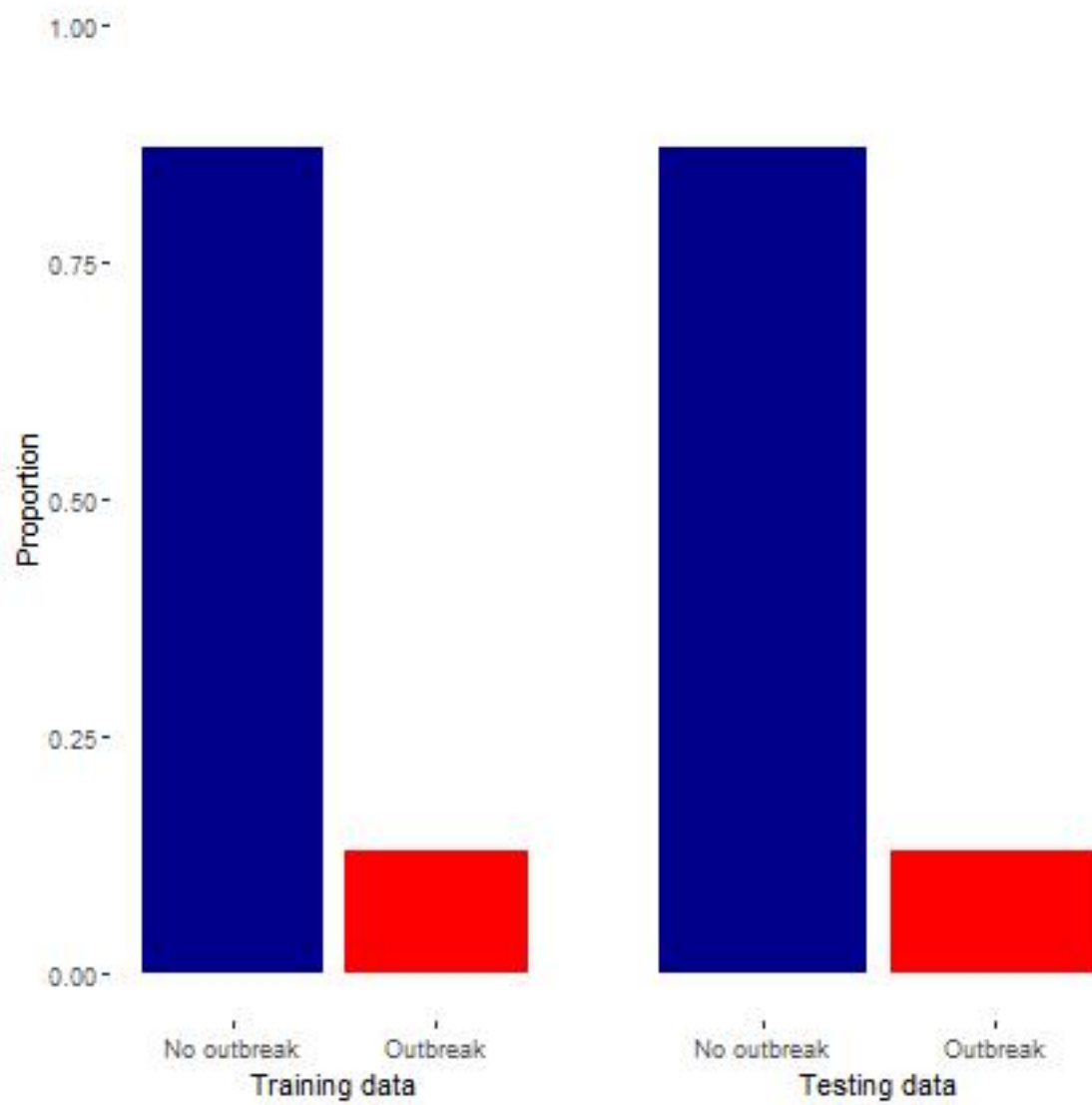


Figure 5.2: Proportion of outbreak in both the training and testing data.

5.2.6 Model validation

All the models were validated using a 10-fold cross-validation. This is done by randomly dividing the training data in to ten mutually exclusive folds of roughly equal size. The model then trained and tested ten times with each time utilizing 9 subsets to train the model and the remaining subset to test it yielding ten performance estimates, such as the model's accuracy. The final performance estimate of the model will be based on the average of the ten estimates as shown in **Eq.5.4**.

$$CV_{PE} = \sum_{i=1}^{10} PE_i \quad (5.4)$$

where PE_i stands for the performance estimate of fold i and CV_{PE} stands for the cross-validation performance estimate.

To enhance the estimated performance, the 10-fold cross validation procedure was repeated five times.

The significance of feature selection in machine learning, particularly in disease prediction, cannot be overstated. In our study, the backward elimination method was instrumental in refining the features, which, as previous research suggests, reduces overfitting and expedites the learning process [123], [124]. This focused approach allows us to leverage the strengths of the models for accurately predicting malaria outbreak in The Gambia.

5.3 Result

The experimentation on each model to predict a malaria outbreak in each district in the Gambia is summarized in this section.

The Pearson correlation among the climatic variables is shown in **Fig 5.3**. Most of the climatic variables are either weakly or moderately correlated except between rainfall and relative humidity with a correlation of 0.7. The correlation between maximum temperature and relative humidity is also slightly strong with a correlation of -0.66 . More importantly, all the results were found to be statistically significant using a significant level of 5% as shown in **Table 5.2**. The 95% confidence interval of the correlation coefficient is also presented in **Table 5.2**.

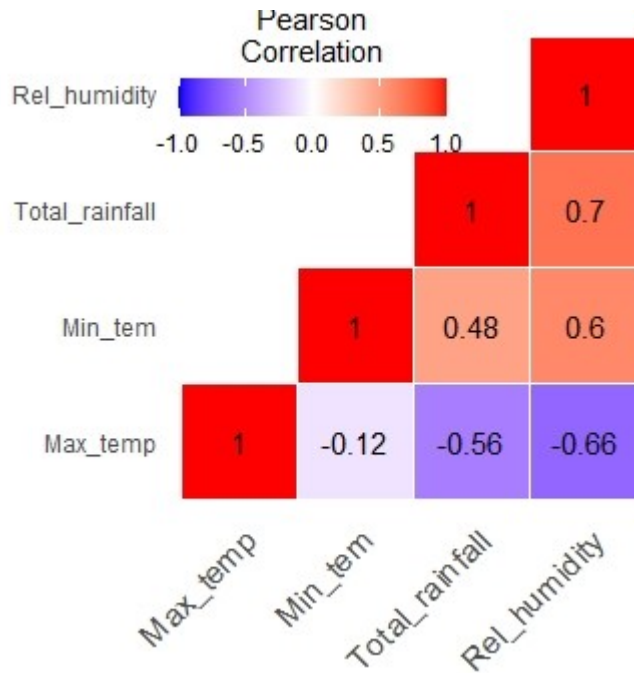


Figure 5.3: The Pearson's correlation among all the climatic variables in the dataset.

Most clinical malaria cases occurred during the end of the rainy season and the two

Climatic Variables	Rainfall	Maximum temperature	Minimum temperature	Relative humidity
Rainfall	(1, 1) $< 2.2e - 16$	(-0.58, -0.54) $< 2.2e - 16$	(0.46, 0.50) $< 2.2e - 16$	(0.68, 0.71) $< 2.2e - 16$
Maximum temperature	(-0.58, -0.54) $< 2.2e - 16$	(1, 1) $< 2.2e - 16$	(-0.15 - 0.09) $2.756e - 16$	(-0.68 - 0.65) $< 2.2e - 16$
Minimum temperature	(0.46, 0.50) $< 2.2e - 16$	(-0.15 - 0.09) $2.756e - 16$	(1, 1) $< 2.2e - 16$	(0.58, 0.62) $< 2.2e - 16$
Relative humidity	(0.68, 0.71) $< 2.2e - 16$	(-0.68 - 0.65) $< 2.2e - 16$	(0.58, 0.62) $< 2.2e - 16$	(1, 1) $< 2.2e - 16$

Table 5.2: **95% confidence interval of the correlation coefficient among the climatic variables with their corresponding p-values.**

months following, roughly from August to December, as depicted in **Fig 5.4**. This period corresponds to when there is abundant greenery and stagnant water in the bushes and on the streets, creating an ideal environment for Anopheles mosquito reproduction. This further emphasizes the role of rainfall in forecasting malaria outbreaks. Except for December, these months also experience the highest average relative humidity. Additionally, **Fig 5.4** illustrates that during these months, the minimum and maximum temperatures ranged from 15.8 to $22.9^{\circ}C$ and 32.25 to $35.33^{\circ}C$, respectively.

We selected features for each model using backward elimination. The features included in the final models are presented in **Table 6.2**. The variables month, district, and population size were fixed to be included in every model, while the climatic variables were selected using the backward elimination method. The only models that chose all of the climatic predictors were XGBoost and ANN. C5.0 DT and svmLinear only failed to include one climatic variable. In contrast, no climate predictor was present in svmRadial.

Table 5.3: Features included in the final predicting models. The tick ($\checkmark$) mark represents the inclusion of a particular variable in the model, whereas if the variable is not included, the space remains blank. M represents months of the calendar year, and Dis represents districts and P is for population size. T_{min} , T_{max} , R , H are minimum and maximum temperatures, total rainfall, and relative humidity, respectively.

Models	Features						
	M	Dis	T_{min}	T_{max}	R	H	P
LR	$\checkmark$	$\checkmark$			$\checkmark$	$\checkmark$	$\checkmark$
KNN	$\checkmark$	$\checkmark$					$\checkmark$
C5.0 DT	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$		$\checkmark$	$\checkmark$
ANN	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$
RF	$\checkmark$	$\checkmark$					$\checkmark$
XGBoost	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$
svmLinear	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$		$\checkmark$
svmRadial	$\checkmark$	$\checkmark$					$\checkmark$

To optimize the performance of each predictive model, hyperparameter tuning was conducted using the grid search method. This approach involves systematically searching through a predefined set of hyperparameter values to identify the combination that yields the optimal model performance.

For instance, the parameter k in the K-Nearest Neighbors (KNN) classifier was explored across the values $K = \{3, 5, 7, 9, 11\}$, with the grid search determining $k = 3$ as the optimal choice. Similar ranges and search procedures were applied to the hyperparameters of other classifiers.

The optimal hyperparameters selected for each predictive model are presented in Table 5.4. These values represent the configurations that resulted in the highest model performance within the specified search space.

The performance measures across various models on both the training and testing datasets as shown in Table 5.5 and 5.6 respectively reveal insightful patterns. In the training set, Random forest (RF) displayed a strong overall predictive capability

Table 5.4: **Optimal parameters selected for each classifier.**

Models	Optimal tuning parameters
LR	none
KNN	k=3
C5.0 DT	trials=20 model=tree winnow=false
ANN	decay=0 size=7
RF	mtry=46.88889
XGBoost	nrounds=200 lambda=0 alpha=0 eta=0.3
svmLinear	C=2.154
svmRadial	C=1000 sigma=100

(accuracy: 99.8%, sensitivity: 100%, specificity: 98.3%). Artificial Neural Network (ANN) and svmRadial also demonstrated high accuracy, sensitivity and specificity, providing a robust performance on the training set.

C5.0 Decision Trees (C5.0 DT) exhibited exceptional accuracy and specificity (99.9% and 99.3%, respectively) in the training set. On the testing set, C5.0 DT maintained consistent performance (accuracy: 93.3%, sensitivity: 96.7%, specificity: 71.6%), showcasing its effectiveness in practical applications.

Extreme Gradient Boosting (XGBoost) demonstrated remarkable consistency, achieving 100% across all metrics on the training set and maintaining high values on the testing set (accuracy: 93.3%, sensitivity: 96.7%, specificity: 71.6%). In summary, C5.0 DT and XGBoost emerge as the top-performing models, offering a well-balanced and robust predictive capacity for malaria outbreak prediction.

The area under the receiver operating characteristic (ROC) curve (AUC) was also

Table 5.5: **Performance measures on the training set.**

Models	Accuracy (%)	Sensitivity (%)	Specificity (%)
LR	88.0	98.9	52.2
KNN	93.6	100	66.9
C5.0 DT	99.9	100	99.3
ANN	96.3	99.9	78.4
RF	99.8	100	98.3
XGBoost	100	100	100
svmLinear	85.9	98.6	47.8
svmRadial	97.9	100	86.2

Table 5.6: **Performance measures on the testing set.**

Models	Accuracy (%)	Sensitivity (%)	Specificity (%)
LR	86.8	98.8	48.6
KNN	89.4	97.7	54.9
C5.0 DT	93.3	96.7	71.6
ANN	89.6	96.0	56.3
RF	91.5	95.4	65.1
XGBoost	93.3	96.7	71.6
svmLinear	84.8	98.4	44.7
svmRadial	91.0	95.0	63.7

used to compare each model’s performance, as shown in **Fig 5.6**. The AUC values of all the models are at least 0.9 except for KNN with 0.89 which is the lowest. XGBoost has the highest AUC value of 0.97 and closely followed by C5.0 DT with an AUC of 0.96.

After considering the performance of all the models with respect to the three popular performance metrics (accuracy, sensitivity and specificity) and the AUC values, we recommend the XGBoost and C5.0 DT model as the best performing models. They have the highest accuracy, specificity and AUC values.

5.4 Discussion

Predictive models for malaria outbreaks must be extremely accurate if they are to have any hope of being helpful in clinical and public health decision-making as well as in planning resources for better and more effective services, especially in areas where medical facilities are scarce. With the use of these models, medical professionals might more effectively plan and deliver treatment and diagnostics to lessen the effects of disease on humans, particularly in endemic areas. In this study, we developed machine learning models to estimate the impact of three climatic variables and some non-climatic variables in predicting malaria outbreak in each district in the Gambia. To the best of our knowledge, this is the first study that attempts to predict malaria outbreak in the Gambia by integrating climatic and some non-climatic variables. However, similar studies have been conducted in different regions around the world to investigate the effect of climatic variables in predicting malaria outbreak [33], [36]. These studies support our finding that demonstrates that machine learning techniques are very useful in predicting malaria outbreak by integrating climatic and non-climatic variables. Similar to this study, all these studies commonly used temperature, rainfall and relative humidity as the climate predictors that can possibly influence malaria outbreaks.

Two distinct datasets from 2013 to 2021 encompassing climatic and non-climatic factors were analyzed in order to validate and compare all nine models. **Fig 5.3** displays the Pearson correlation among the climatic variables. **Table 5.5** and **5.6** displays the sensitivity, specificity, and accuracy of each classifier’s performance. Despite

the fact that each machine learning technique has unique strengths and weaknesses depending on the application type and size of the dataset, all of the models produced promising results in terms of accuracy, sensitivity, specificity as well as AUC values as shown in the result section. We selected XGBoost and C5.0 DT as the best performing models.

We also applied feature selection algorithm technique to return the overall best model with just the relevant features. The non-climatic features (District, population size and month) were forced into all the nine models while the climatic variables were selected by the feature selection algorithm. The variable month was included in all of the predictive models in order to forecast malaria cases for every month in The Gambia. This is because malaria transmission in The Gambia highly varies with respect to the month. It was reported in [125] that The Gambia records its highest clinical cases and fatalities in September and November. This report supports our findings, presented in **Fig.5.4**, showing November recording the highest number of malaria cases in The Gambia. The variable District was also fixed to enable us predict malaria cases for each district in The Gambia. From the feature selection results, all the models except for KNN, RF and svmRadial have included at least two climatic features showing why they can be a crucial variables in predicting malaria outbreaks. Each of the climatic variable appeared in four models. The minimum and maximum temperature happened to be selected by the same models. One of the two best models, XGBoost selected all the climatic variables while the second one, C5.0 DT selected all except rainfall. This has further strengthened the significance of these climatic variable in predicting outbreak. Furthermore, these three climatic

factors have been linked to the spread of malaria and the development of the anopheles mosquito, according to numerous research [31], [32], [96].

In conclusion, this study examined the significance of machine learning methods for predicting malaria outbreaks in The Gambia. It is crucial to anticipate malaria outbreaks in order to lower the disease’s morbidity and fatality rates in The Gambia. And more importantly, the accurate forecasting of malaria outbreak can give public, health authorities and clinical health services the knowledge they need and the required steps to conduct targeted malaria control strategies that make efficient use of scarce resources to reduce loss of lives.

We employed a number of machine learning techniques to forecast malaria outbreaks in each district in The Gambia. Using historical meteorological data, population size, and malaria case data spanning 9 years (2013–2021), all models were trained, validated, and evaluated. The findings indicated that a combination of meteorological and non-climatic factors like population size, month are significant in predicting when malaria epidemics would occur in a particular region or district in The Gambia. Moreover, we recommended XGBoost and C5.0 DT as the best models for prediction based on reasons highlighted in the result section.

This research can be enhanced in the future by utilizing the hybridized ensemble approach of machine learning. The final model will be more reliable as a result. The research can also be further strengthened by including other elements like treated mosquito nets, indoor residual spray coverage, and other pertinent predictors.

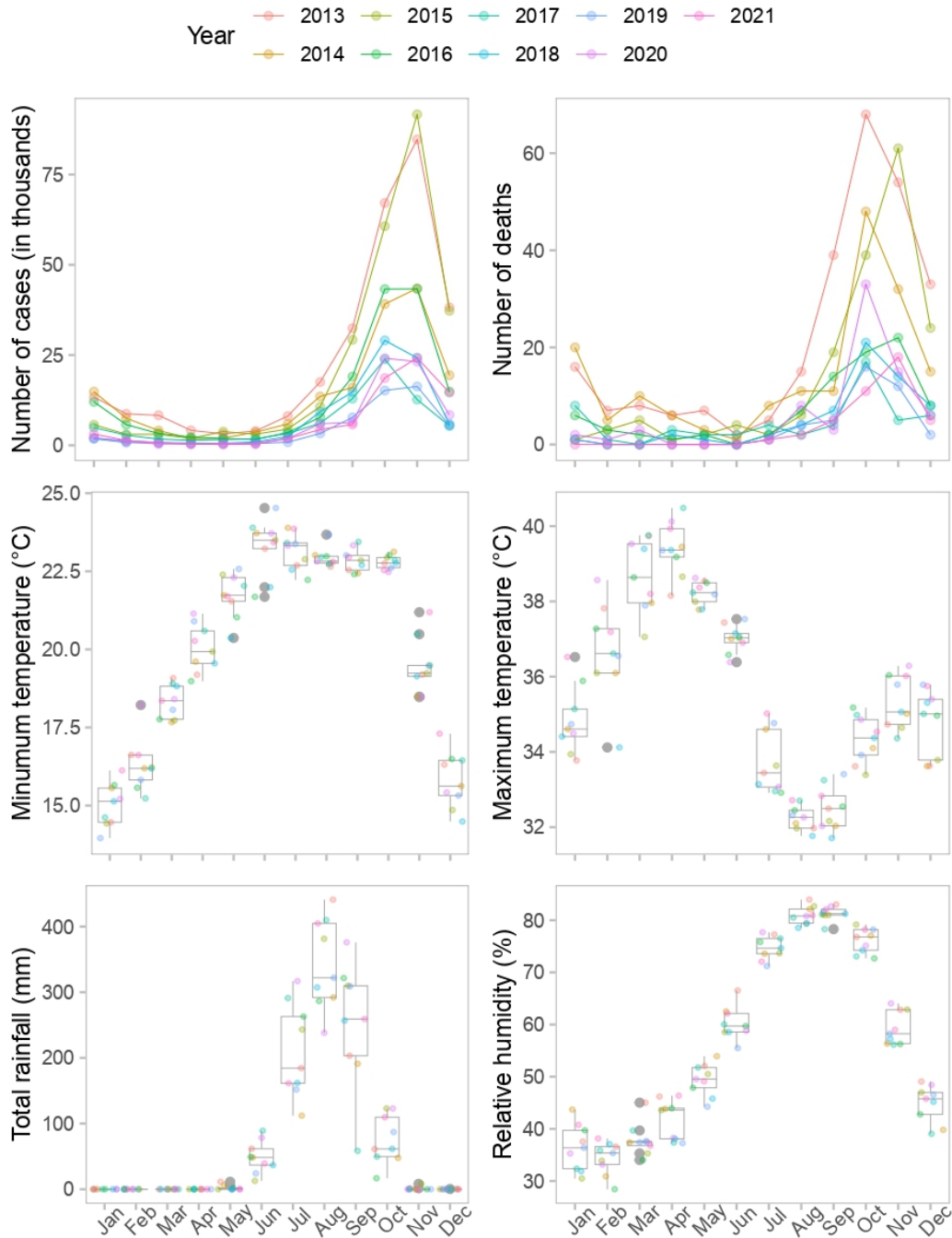
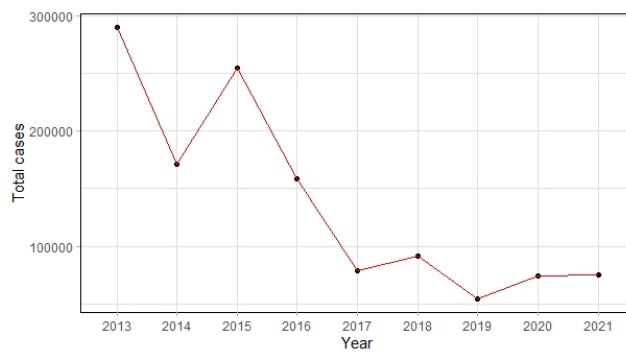
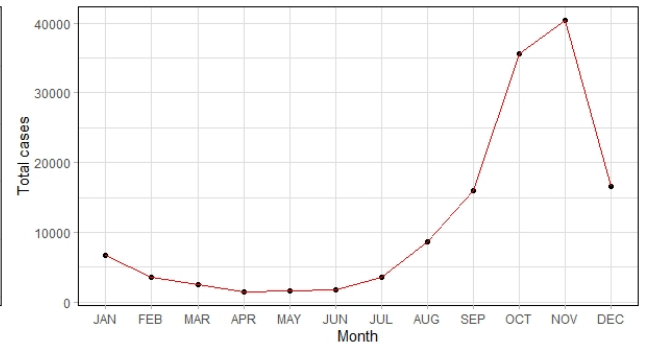


Figure 5.4: **Yearly and Monthly Variations of Malaria Cases, Deaths, and Climatic Factors.** The upper panels display the yearly fluctuations in malaria cases and deaths, with each color-coded line corresponding to a specific year. The lower four panels illustrate the variability in climatic factors. Boxplots represent monthly data, while individual data points for each year are represented by colored dots. Temperature measurements are provided in Celsius ($^{\circ}\text{C}$), rainfall in millimeters (mm), and relative humidity in percentages (%).



(a) Trend of malaria cases over the years



(b) Average cases per month over the years

Figure 5.5: **Temporal Trends in Malaria Cases from 2013 to 2021.** A noteworthy decline in malaria cases is observed over the years, with the highest incidence in 2013 and the lowest in 2019. A slight increase in cases is seen post-2019. Peak malaria cases are typically recorded from August to December, with November and October being the most affected months. These patterns align with the latter half of the rainy season.

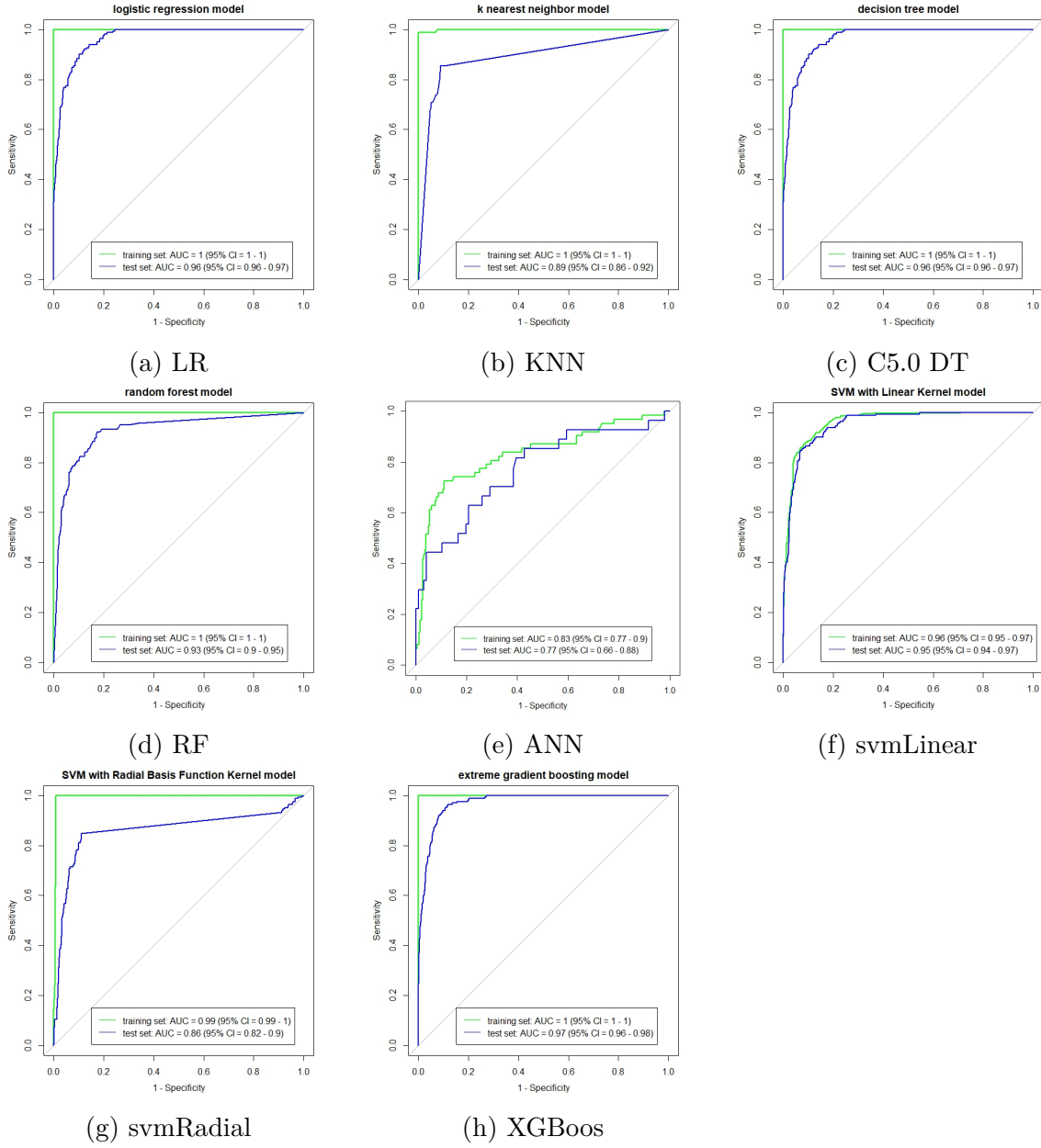


Figure 5.6: This plot shows the Area Under the Receiver Operating Characteristic (ROC), (AUC) of all the learning algorithms. The green and the blue curve shows the ROC curve using the training and testing data respectively. XGBoost has the highest AUC value of 0.97.

6 Machine Learning for Predicting Malaria Incidence in each district in The Gambia

6.1 Objective

The objective of this chapter is to address the dire challenge of malaria incidence in The Gambia, particularly in light of increasing resistance to conventional control measures and existing infrastructural limitations. To tackle this issue, we employ the quantile random forest (QRF) regression model to develop a predictive model for malaria incidence across different districts in The Gambia. By utilizing both climatic and non-climatic data, our goal is to establish a robust early warning system. This tool is anticipated to bolster preparedness, facilitate timely and strategic interventions, and enhance resource allocation, thereby making a substantial contribution to proactive malaria control efforts in The Gambia.

6.2 Method

6.2.1 Study area

See **Study area** in the previous chapter.

6.2.2 Dataset used

This chapter utilizes the same clinical and climatic dataset as in the previous chapter. For more details, see **Dataset used** and **Data Preprocessing** in the previous chapter.

6.2.3 Machine Learning Model: Quantile Random Forest (QRF)

In this paper, we employed the QRF algorithm to predict the number of malaria cases across districts in The Gambia. The QRF algorithm is an extension of the Random Forest (RF) method, designed to provide robust predictions, especially in the presence of outliers [129].

For evaluation, all the models underwent training on 70% of the data, with the remaining 30% earmarked for validation. Overfitting mitigation was achieved using a rigorous 10-fold cross-validation process. Our evaluation metrics encompassed the coefficient of determination R^2 , root mean square error (RMSE), and mean absolute error (MAE).

The significance of feature selection in machine learning, particularly in disease

prediction, cannot be overstated. In our study, the backward elimination method was instrumental in refining the features, which, as previous research suggests, reduces overfitting and expedites the learning process [123], [124]. This focused approach allows us to leverage the strengths of the QRF model for accurate malaria case predictions in The Gambia.

6.3 Result

6.3.1 Data Summary

In this section, we summarise the key findings of our study. Table 5.1 in **Data Preprocessing** provides a comprehensive list of all variables present in the dataset.

As substantiated by previous research [126], malaria transmission in Gambia typically begins in July or August and continues through December, with minimal or no clinical cases observed during the dry season. The Upper River Region (URR), located in the eastern part of Gambia, is particularly noteworthy because of its high prevalence of malaria. This region is characterised by distinct seasonal malaria transmission, low vector densities, and high vector survival rates (parous rates range from 81% to 91% compared to 27–71% in other regions) [127].

Figures 5.5 and **5.4** show a consistent pattern: the highest average number of cases was recorded from August to December. Specifically, October and November had the highest average number of malaria cases throughout the nine-year period, followed by August and December. The significant role of the 'month' variable in predicting

malaria cases in The Gambia echoes the findings in [126]. This period, characterised by abundant greenery and stagnant water bodies, serves as an ideal breeding ground for *Anopheles* mosquitoes. In addition, **Fig.5.4** illustrates the correlation between rainfall patterns in Gambia and the month of the year, identifying rainfall as a critical predictor of malaria cases. Despite the relative absence of rainfall in most months, the monthly rainfall variability aligns with the seasonality of malaria cases, underscoring the crucial role of rainfall in predicting malaria cases in Gambia.

6.3.2 Analysing the Influence of Climatic Variables

This section delves deeper into the complex relationship between climatic variables and malaria cases reported in Gambia from 2013 to 2021. Climate factors are pivotal to public health, especially in vector-borne diseases, such as malaria. A thorough understanding of how key climatic variables such as temperature, rainfall, and relative humidity influence malaria incidence is imperative for formulating effective strategies to mitigate malaria in Gambia.

Several studies [31], [32], [96] have identified significant correlations between these climatic variables and both the spread of malaria and life cycle of *Anopheles* mosquitoes. For instance, [31] posited that certain climatic conditions are conducive to the proliferation of the *Anopheles* mosquito and *Plasmodium* parasite, both of which are intrinsically linked to malaria prevalence. Furthermore, [96] highlighted that rainfall, temperature, and relative humidity were the primary climatic variables influencing the development and maturation of *Plasmodium* parasites. As such, any alterations in these factors will invariably affect mosquito ecology.

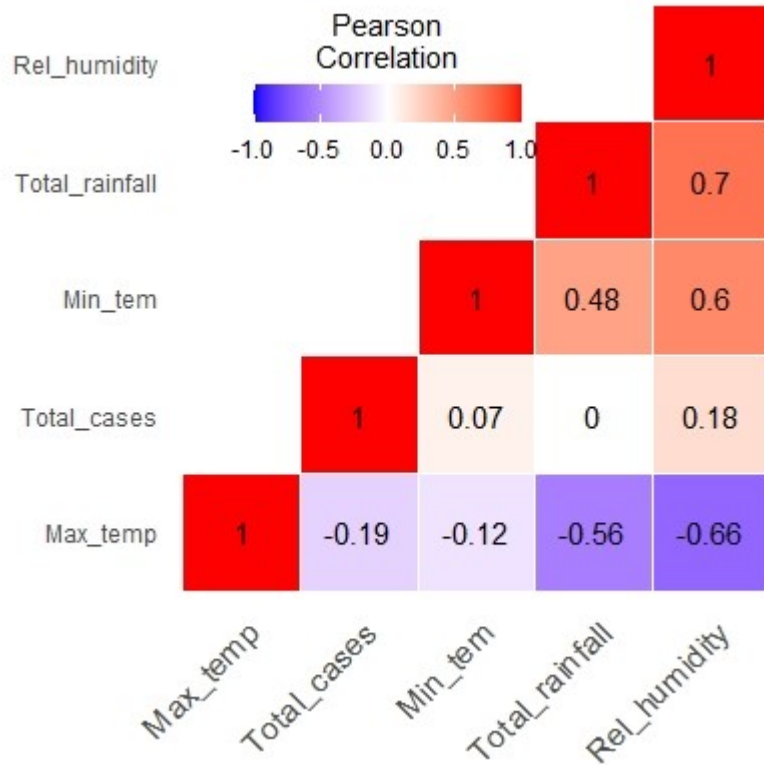


Figure 6.1: **Pearson correlation between the climatic variables and total malaria cases**

[32] emphasises rainfall and temperature as the key elements influencing malaria transmission, noting that these variables also affect the lifespan of adult mosquitoes. Higher temperatures accelerate the development of both parasites and vectors [130]. However, malaria transmission efficiency is subject to various factors including vertebrate hosts, parasites, and vectors [131]. Interestingly, the peak of malaria transmission in The Gambia occurs during the rainy season and the subsequent couple of months, predominantly from June to October or November [30]. These periods corresponded to high relative humidity and the highest minimum temperatures, as shown in **Fig.5.4**, further underscoring the influence of these climatic conditions on the prevalence of the disease.

However, from the Pearson correlation plot shown in **Fig.6.1**, the total malaria cases is weakly correlated with the climatic variables, where as the climatic variables are moderately correlated to each other with rainfall and relative humidity being the most correlated.

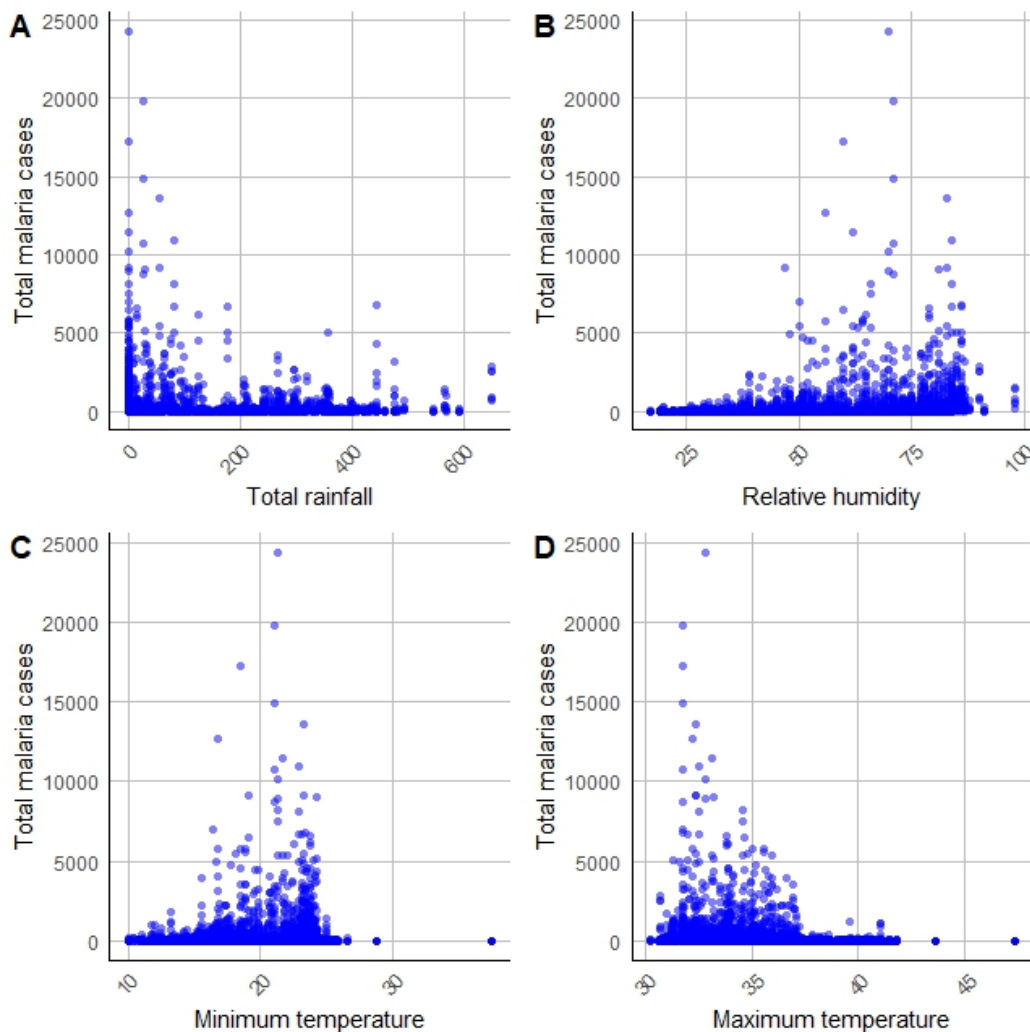


Figure 6.2: **Relationship between environmental factors and malaria cases.** Each subplot depicts the relationship between a specific environmental factor (total rainfall, relative humidity, minimum temperature, and maximum temperature) and the total number of malaria cases.

From Fig.6.2, we can see the various combination of these climatic features that give rise to high malaria cases in The Gambia. Most high malaria cases were recorded

between 0 to 200mm of rainfall, between 60 to 80% of relative humidity, minimum temperature between 18 to 22 degree Celsius and maximum temperature of 32 to 36 degree Celsius.

6.3.3 Prediction Model

In this study, we employed the quantile random forest (QRF) model to examine the intricate relationship between selected features and malaria incidence in Gambia. **Fig.5.4** clearly delineates the monthly fluctuations of the three pivotal climatic variables, independent of the year. **Table 6.2** provides a breakdown of the features incorporated into the QRF model with district, month, and population size as significant predictors.

Table 6.1: Performance Metrics for Training and Testing Datasets

Performance Metric	Dataset	
	Training	Testing
R^2	0.961	0.642
RMSE	237.856	117.139
MAE	37.932	50.116

The performance metrics of the QRF model based on the training and testing datasets are shown in **Table 6.1**. In the training phase, the QRF model achieved an impressive R^2 value of 0.961, with a Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) of 237.856 and 37.932, respectively. However, it is crucial to note the potential for overfitting given the model's outstanding performance on the training dataset.

In the testing phase, the model's performance is notably lower, with an R^2 value

of 0.642, an RMSE of 117.139, and an MAE of 50.116. This discrepancy suggests that the model may have learned the training data too well, capturing noise and specific patterns that don't generalize well to new, unseen data. These results raise concerns about the model's ability to generalize effectively.

Model	Features						
	<i>Dis</i>	<i>M</i>	<i>T_{min}</i>	<i>T_{max}</i>	<i>R</i>	<i>H</i>	<i>P</i>
QRF	✓	✓		✓		✓	✓

Table 6.2: **Selected features in the QRF model.** A variable selected for the model is shown with a tick (✓) mark, whereas if the variable is not included, the space remains blank. *M*, *Dis*, and *P* represent months of the year, districts in the Gambia, and population for each district respectively. *T_{min}*, *T_{max}*, *R*, *H* are minimum and maximum temperatures, total rainfall, and relative humidity, respectively.

The QRF model has identified crucial features for accurate predictions, including district (*Dis*), month (*M*), Maximum Temperature (*T_{max}*), Relative Humidity (*H*), and population (*P*), as highlighted in **Table 6.2**. These selections underscore the significant impact of environmental and demographic factors in achieving precise malaria case predictions. By effectively incorporating key climatic and demographic predictors, such as maximum temperature, relative humidity, and population, the QRF model establishes a robust methodology for forecasting malaria prevalence in Gambia.

In alignment with the methodology employed in the previous chapter, the parameter tuning technique remains consistent in this study. The grid search method, a systematic exploration of a predefined set of hyperparameter values, is once again employed to optimize the performance of the predictive models. For instance, the QRF model's optimal tuning parameter, *mtry* = 33.77778, was determined through

an exhaustive search within a specified range. This parameter tuning step plays a crucial role in enhancing the model's predictive accuracy, ensuring that it is finely calibrated to the unique characteristics of the dataset.

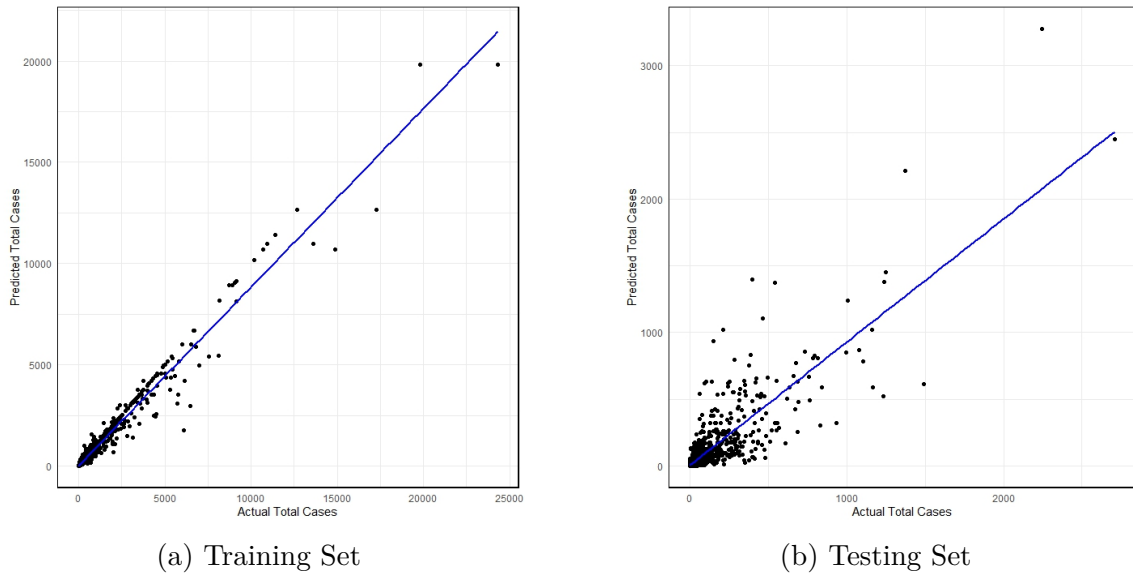


Figure 6.3: Parity Plots for Training and Testing Sets. Comparison of actual and predicted total malaria cases. The blue regression line indicates a close alignment between the actual and predicted values.

Fig.6.3 displays the Parity Plots for both the training and testing Sets. These plots showcase the comparison between the actual and predicted total malaria cases. The inclusion of the regression line in blue provides insight into the overall trend, indicating a remarkably close alignment between the actual and predicted values especially in the training set. The observed divergence in the Parity Plot for the testing set suggests potential overfitting issues, indicating that the model's performance, while closely aligned in the training set, may encounter challenges when applied to new, unseen data due to its excessive tailoring to the training set characteristics.

6.4 Discussion

The utilization of machine learning (ML) models, as well as time series models like the autoregressive integrated moving average (ARIMA) and seasonal autoregressive integrated moving average (SARIMA), has become pivotal in infectious disease prediction including malaria. In a study conducted by Khan et al. [132], multiple ML techniques were employed to predict malaria outbreaks across various districts in The Gambia. Their findings underscored the effectiveness of ML models in utilizing both climatic and some non-climatic data for accurate predictions.

Similarly, Ebhuoma et al. [133] tackled the prevalence forecasting of malaria in KwaZulu-Natal, South Africa, employing three distinct models. Among these, the SARIMA model emerged as the most fitting, demonstrating its potential in capturing the intricacies of malaria incidence trends. The model's predictive capabilities were corroborated by its close alignment with actual malaria cases recorded in 2014.

Another study conducted in Hefei, China, by Zhai et al. [134] used both monthly malaria cases and climatic data spanning over two decades (from January 1990 to December 2011). Through the application of logistic regression models integrated with time series seasonal decomposition, the researchers investigated the influence of climatic and seasonal conditions on malaria outbreaks. The empirical model developed, grounded in disease surveillance and climatic factors, exhibited promising potential for malaria control and prevention endeavors.

In this chapter, utilizing data spanning from January 2013 to December 2021, the Quantile Random Forest (QRF) model emerged as the primary predictive tool for fore-

casting malaria cases across diverse districts in The Gambia. By integrating a range of climatic and non-climatic factors, our study showcases the robustness of the QRF model. Its feature selection process highlighted critical variables, including District, Month, maximum temperature, relative humidity, and population size. Notably, the exclusion of rainfall, despite its established relevance in malaria transmission dynamics, sheds light on the model's unique approach.

The observed correlation between the number of malaria cases and the month further emphasizes the influence of seasonal variations on disease incidence. Months such as September, October, and November registered notably higher case counts, while the period from January to July saw relatively fewer cases. This can be seen in **Fig.2.13** and **5.4**. This stark variability in cases contributes to the challenges of modeling and forecasting, particularly when extrapolating to new data. The inclusion of "population size" for each district adds another layer of complexity, with districts varying significantly in demographic composition. This variance in population size undoubtedly influences the incidence of malaria cases, as reflected in the dataset.

Overall, the QRF model stands out as a robust tool for malaria prediction in The Gambia. Its demonstrated performance on both training and testing datasets, especially when confronted with untried data, solidifies its position as a reliable predictive model. This efficacy positions the QRF model as a valuable resource for public health professionals and stakeholders in devising proactive strategies to combat and ultimately eradicate malaria. Its role as an early-warning system and resource allocation tool holds immense potential in optimizing the distribution of crucial interventions, including insecticide-treated nets, residual sprays, and antimalarial medications, es-

pecially in areas far from medical facilities. The predictive power of the QRF model is poised to play a pivotal role in the ongoing battle against malaria in The Gambia.

To the best of our knowledge, this is the first study that aims at predicting malaria cases in the Gambia using climatic variables. To predict malaria cases in each district of The Gambia, we employed a machine learning technique (QRF model) using climatic data variables, month and population size as predictors. The model was trained, validated and assessed. Our research has shown that climatic characteristics are a significant factor in predicting malaria cases in each district in The Gambia. From the performance, it is clear that the QRF model demonstrated remarkable proficiency in capturing the intricate relationship between these factors and malaria incidence.

Future research should take into account forecasting malaria cases for a specific month or a specific district in the Gambia due to the high variability of total cases as a result of the differences in monthly transmission of malaria in the Gambia and the differences in population size for each district in the Gambia. Other non-climatic factors including treated mosquito nets, indoor residual spray coverage, and other relevant predictors should also be considered in future studies as well.

7 Integrating FAMD and Imputation methods in High Dimensional Mixed-Data Type

7.1 Objective

The focus of this chapter is on addressing the burning challenge of missing data in high-dimensional mixed-type datasets. The objective is to impute missing values in high dimensional mixed-type datasets. We first apply factor analysis of mixed data (FAMD) to reduce the dimension of the dataset and then applied and compared different imputation method for mixed-data type. Three techniques are compared: the K-Nearest Neighbors (KNN) with a specialized distance measure proposed in [45] to impute the missing values in high dimensional dataset, Random forest method and MICE imputation.

7.2 Methodology

In this section, we present the approach to imputing missing values in high-dimensional mixed-type datasets. Our method hinges on first applying Factor Analysis for Mixed Data (FAMD) to reduce the dimension of the dataset and then apply the imputation techniques. A critical consideration for the nearest neighbor imputation proposed in [45] lies in the definition of an effective distance measure. The performance of nearest neighbors imputation depends crucially on the distance used to find these neighbors. Most distance measures are designed to handle only continuous data.

7.2.1 Factor Analysis for Mixed Data (FAMD)

FAMD serves as a pivotal tool for dimensionality reduction in datasets encompassing both continuous and categorical variables. By leveraging the underlying structure of the data, FAMD enables the extraction of relevant features while mitigating information loss.

Consider a dataset X with p features, denoted as $\{x_1, x_2, \dots, x_p\}$ where x_{ip} denotes the i th observation of the p th feature. Without loss of generality, we assume that the first q features have complete information, while the remaining $p - q$ features contain some missing values. This is because any datasets can be rearranged so that the first q features have no missing values while the remaining features contain a few missing values. This configuration allows us to focus on imputing missing values in the latter subset.

To enhance the imputation process, we begin by reducing the dimensionality of

the features in X with complete information. This set of features is represented as A , comprising the first q features while the remaining $p - q$ features are represented as B . We apply FAMD to transform A into a lower-dimensional space, denoted as A' , while retaining critical information [135]. This step is pivotal in mitigating the challenges posed by high-dimensional datasets, as it strikes a balance between dimensionality reduction and information preservation.

FAMD accomplishes this by identifying the underlying structure and patterns in the data. It does so by deriving a set of orthogonal axes, or principal components, that capture the directions of maximum variance in A . These components provide a reduced set of features that retain the essential information required for accurate analyses. After applying FAMD on A , we combine the transformed A which is A' with B , which contains the missing values. This give rise to a new reduced dataset, R . Hence we apply the imputation techniques on R instead of X .

7.2.2 k-Nearest Neighbors (kNN) Imputation

Subsequent to the dimensionality reduction accomplished by FAMD, we employ the imputation techniques in the reduced dataset R . The kNN imputation method identifies neighboring data points to a given observation based on a customized similarity metric tailored for mixed-type data. The performance of nearest neighbors imputation depends crucially on the distance used to find these neighbors.

Distance measure

In this section, we introduce the weighted distance metric for handling mixed-type data in knn method. Let $\mathcal{R} = (R_{is})$ denote the $n \times (u + v)$ data matrix, comprising u continuous and v categorical covariates, represented as $\mathcal{R} = (X, Y)$. Here, $X = (x_{ig})$ for $g = 1, \dots, u$, signifies the g -th continuous covariate with x_{ig} representing the i -th observation. Similarly, $Y = (y_{il})$ for $l = 1, \dots, v$, denotes the l -th categorical covariate with y_{il} denoting the i -th observation.

To incorporate missing values into the computation, we define the binary matrix $\mathcal{O} = (o_{is})$, where $o_{is} = 1$ if R_{is} is observed, and $o_{is} = 0$ otherwise.

For categorical variables with k_l possible values ($c_l \in \{1, \dots, k_l\}$), each y_{il} is transformed into a binary vector $\mathbf{y}_{il} = (y_{il1}, \dots, y_{ilk_s})$, where $y_{ilr} = 1$ if $Y_{il} = r$. This transformation allows us to represent the categorical covariates effectively.

Considering missing values, the distance between the i -th and j -th rows is given by:

$$d_{q,s}(R_i, R_j) = \left(\beta_1 \sum_{g=1}^u |x_{ig} - x_{jg}|^q I(o_{ig} = 1) I(o_{jg} = 1) C(\delta_{sg}) + \beta_2 \sum_{l=1}^v \sum_{c=1}^{k_l} |y_{ilc} - y_{jlc}|^q I(o_{il} = 1) I(o_{jl} = 1) C(\delta_{sl}) \right)^{1/q} \quad (7.1)$$

Here, $I(\cdot)$ denotes the indicator function, and β_1 and β_2 are weights, subject to the constraint $\beta_1 + \beta_2 = 1$. For $q = 1$, the distance corresponds to the L1 distance, while for $q = 2$, it represents the L2 distance.

The formulation presented above closely follows the methodology outlined in [45],

as it provides a robust framework for measuring distances in mixed-type data, effectively accommodating both continuous and categorical variables.

Weights selection

Central to the distance measure is the selection of weights, denoted as β_1 and β_2 . These weights govern the relative influence of continuous and categorical variables in the distance computation. A higher value of β_1 places greater emphasis on continuous attributes, while a higher value of β_2 highlights the significance of categorical attributes.

To ensure an optimal choice of weights, we employ cross-validation. This process helps fine-tune β_1 and β_2 to align with the specific characteristics of the dataset and research objectives. The sum of β_1 and β_2 is normalized to 1, ensuring a balanced allocation of influence between the two variable types. This normalization maintains the integrity of the distance measure.

Weighted imputation by the nearest neighbors

The weighted imputation by nearest neighbors method employed in this chapter closely parallels that outlined in [45]. This approach utilizes a kernel-weighted nearest neighbors scheme, optimizing the imputation process by assigning varying weights to observations based on their proximity. Specifically, Gaussian kernels are employed, and the choice of tuning parameters is determined via cross-validation. For both categorical and continuous variables, the imputation process follows a similar framework, utilizing weighted estimators to achieve accurate imputed values. Notably, our

method incorporates an additional step of applying Factor Analysis of Mixed Data (FAMD) before employing K-Nearest Neighbors (KNN) imputation, distinguishing it from the approach presented in [45].

7.2.3 Random Forest Imputation

Initially proposed by Breiman [59] as an extension of classification and regression trees (CART), Random Forests have been widely employed in various domains, including missing data imputation [101]. This approach is versatile, accommodating both categorical and continuous data, even in scenarios involving a high number of predictors. It employs an iterative strategy, commencing with a basic imputation technique such as mean imputation and progressively refining the imputed data with each iteration. We applied this imputation technique to impute missing values after FAMD for dimensionality reduction just like the knn approach.

7.2.4 Multiple Imputation by Chained Equations (MICE)

The Multiple Imputation by Chained Equations (MICE) was also compared with the other two technique after dimensionality reduction. MICE is a powerful imputation technique that addresses missing data by iteratively imputing incomplete variables based on conditional distributions. Each incomplete variable is imputed using a regression model that takes into account the observed variables. This process is repeated multiple times to create multiple imputed datasets. MICE has shown effectiveness in handling mixed-type data, making it a relevant method for our comparative evaluation.

7.2.5 Performance measures

A reliable imputation technique strives to provide estimates that closely align with the actual true values. To assess the effectiveness of our proposed imputation methods, we employ two primary performance metrics: Mean Squared Imputation Error (MSIE) and Mean Absolute Imputation Error (MAIE).

MSIE measures the difference between the imputed values in the $p - q$ features after introducing NA's and the original true values in the same subset before the introduction of NA's. Denoting the imputed NA's as $\hat{x}_{im}$ and the true values as x_{im} , the MSIE (as depicted in **Eq.7.2**) serves as the benchmark for evaluating performance.

Similarly, MAIE quantifies the average absolute difference between imputed and true values, providing an alternative perspective on imputation accuracy. MAIE can be calculated using the formula in **Eq.7.3**.

We conduct a comparative analysis of the MSIE and MAIE for the three imputation methods, missForest [101] from Random Forest, MICE [136], and the weighted knn imputation method proposed in [45].

$$\text{MSIE} = \sqrt{\frac{1}{n(M)} \sum_{x_{im} \in \mathcal{M}} (\hat{x}_{im} - x_{im})^2} \quad (7.2)$$

$$\text{MAIE} = \frac{1}{n(M)} \sum_{x_{im} \in \mathcal{M}} |\hat{x}_{im} - x_{im}| \quad (7.3)$$

7.3 Simulation Setup

We conducted extensive simulations to evaluate the performance of our proposed method. We compare the effectiveness of three distinct imputation techniques: kNN imputation, random forest, and the standard MICE method across various missing data patterns.

To determine the optimal value of q in the knn imputation method, we referred to the study in [45], which conducted simulations comparing $q = 1$ and $q = 2$. Their results strongly favored $q = 2$ as the superior choice. Consequently, we have adopted $q = 2$ for our experiments for consistency with their findings.

To emulate real-world scenarios, we generated $S = 500$ samples, each comprising $n = 20$ observations for $p = 50$ covariates. These samples were drawn from a multivariate normal distribution with a mean of 0 and a covariance matrix Σ characterized by an autoregressive structure of type AR(1) with $\rho = 0.9$.

7.3.1 Varying Missing Percentages in the $p - q$ features

We started by randomly selecting 25 variables (50% of the total) and transformed them into nominal scale by setting appropriate cut points for the continuous covariates, creating meaningful categories. Subsequently, we divided the dataset into two portions: the first containing q features (50% of the data) and the remaining containing $p - q$ features (50% of the data). Next, we introduced missing values randomly in the second portion, simulating different missingness patterns. We considered missing rates of 10% and 20%. For instance, a 10% missing rate indicates that 10% of the

entries in the second portion are missing.

7.3.2 Variations in Data Characteristics

To explore the influence of data characteristics on imputation effectiveness, we varied both the dimensionality and composition of the dataset. Specifically, we considered scenarios with $p = 20$, $p = 30$, and $p = 50$ covariates, while keeping $n = 20$ constant. In each sample, 50% of the variables were initially converted to categorical format, with the remaining variables retained as continuous.

Additionally, we extended our investigation to explore the impact of different compositions of categorical and continuous variables. In previous experiments, we consistently transformed 50% of the variables into categorical format. To gain a more detailed understanding, we now explore scenarios where 60%, 70%, and 80% of the variables are converted to categorical. This systematic variation provides valuable insights into how the proportion of categorical features influences imputation effectiveness.

Note:In both cases, the data was divided into two portions, where the first portion (50%) was free of missing values and missing values introduced into the remaining portion (50%).

The simulation studies outlined here were conducted based on the methodology outlined in the Proposed Method section. This involved applying Factor Analysis of Mixed Data (FAMD) on the first portion (A) containing q features with no missing values, followed by the imputation methods on the combined data from the second portion (B) containing $p - q$ features with missing values and the reduced portion

(A') from FAMD. This resulted in a new dataset $R = [B, A']$.

7.4 Result and discussion

We present the result of our simulation study in this section.

Table 7.1: **Effects of different missing percentages, for $p=20$ covariates, $\rho = 0.9$ and percentage of categorical variables=50%.**

Methods		Missing Percentage	
		10%	20%
Knn method	MSIE	1.1257	1.2014
	MAIE	1.3107	1.3119
Random forest	MSIE	0.1003724	0.215032
	MAIE	0.1018297	0.2170528
MICE method	MSIE	0.08733446	0.1299649
	MAIE	0.08761625	0.1260317

Table 7.2: **Effects of different missing percentages, for $p=30$ covariates, $\rho = 0.9$ and percentage of categorical variables=50%.**

Methods		Missing Percentage	
		10%	20%
Knn method	MSIE	1.00245	1.13442
	MAIE	1.17620	1.21780
Random forest	MSIE	0.1362539	0.2456587
	MAIE	0.1395311	0.2508075
MICE method	MSIE	0.09744402	0.2649272
	MAIE	0.1030355	0.2574738

The presented tables depict the performance of the three distinct imputation methods, namely Knn, Random Forest, and MICE, under varying configurations of covariate dimensions (p) and missing data percentages. The analysis provides insights into the efficacy of these methods for imputing missing values in high dimensional mixed

Table 7.3: **Effects of different missing percentages, for $p=50$ covariates, $\rho = 0.9$ and percentage of categorical variables=50%.**

Methods		Missing percentage	
		10%	20%
Knn method	MSIE	1.41238	1.32784
	MAIE	1.55409	1.56406
Random forest	MSIE	0.1338755	0.2553096
	MAIE	0.1380639	0.262612
MICE method	MSIE	0.1037395	0.2214063
	MAIE	0.09774569	0.1991495

datasets set apart by different dimensions and levels of missingness.

Beginning with the case of $p = 20$ covariates, it can be seen that the Knn method exhibits relatively higher Mean Squared Imputation Error (MSIE) and Mean Absolute Imputation Error (MAIE) values compared to both Random Forest and MICE. This trend persists across different missing percentages, suggesting that the Knn with a specialized distance measure used may not be the most suitable imputation technique for datasets with a relatively lower number of covariates compared to the other two.

Conversely, the Random Forest method demonstrates superior imputation performance with lower error rates across all missing percentages for $p = 20$. This suggests that Random Forest, utilizing its ensemble learning capabilities, is adept at capturing complex relationships within the data and consequently provides more accurate imputations.

For $p = 30$ covariates, a similar trend is observed. The Knn method again yields higher imputation errors, in both the MSIE and MAIE, compared to Random Forest and MICE. For a moderately high dimensional data, the knn method also performs the worst. However, the most obvious observation is the consistent superiority of

the MICE method across all configurations of p and missing percentages. Regardless of the covariate dimension and the proportion of missing data, MICE consistently outperforms both Knn and Random Forest in terms of imputation accuracy. This is indicative of MICE's robustness and adaptability in scenarios with varying data characteristics.

As the dimensionality of the covariates increases to $p = 50$, a similar pattern persists. The Knn method exhibits higher imputation errors, indicating its limited effectiveness in high-dimensional settings. Random Forest, while performing better than Knn, still falls short of the imputation accuracy achieved by MICE.

In summary, the findings highlight how crucial it is to choose an imputation strategy that is suitable for the particular features of the dataset. Random Forest is an excellent option for datasets with comparatively fewer covariates because it shows competitive imputation accuracy. But the MICE technique regularly outperforms both Knn and Random Forest, yielding the highest accurate imputations, in scenarios with more covariates. The comparison of these three imputation methods lies on the preceding application of Factor Analysis of Mixed Data (FAMD) to reduce the dataset's dimensionality. This preliminary step, while not commonly employed in imputation, significantly influenced the observed results. Therefore, without prior FAMD, it remains uncertain whether MICE would outperform Knn and Random Forest, particularly in scenarios involving higher covariate dimensions.

With the use of FAMD, based on our results, the MICE approach is recommended for real-world scenarios where imputation accuracy and computational efficiency is crucial, especially for high-dimensional datasets. It is a reliable option for imputing

missing values in a variety of real-world settings due to its constant performance across different data setups.

Incorporating FAMD into the method, our findings underscore MICE as the preferred approach for real-world scenarios prioritizing imputation accuracy and computational efficiency, particularly in high-dimensional datasets. Its consistent performance across diverse data configurations positions MICE as a reliable choice for missing value imputation in various real-world applications.

However, it's important to acknowledge some limitations. The evaluation was conducted using simulated data, and the performance may vary when applied to real-world datasets. Additionally, only a specific correlation measure, $\rho = 0.9$ was utilized; exploring other correlation measures could provide further insights into the methods' performance.

8 Conclusion and Future

Directions

8.1 Conclusion

The studies on predicting breast cancer recurrence, as well as modeling malaria dynamics, have represented significant strides in the field of predictive healthcare. Leveraging datasets from King Fahd Specialist Hospital in Dammam and The Gambia, these studies share a unified objective: to develop accurate prognostic tools for enhancing patient outcomes.

The logistic regression model showcased commendable predictive power in the breast cancer studies, achieving an AUC of 76% for both training and test sets. This model identified critical determinants such as chemotherapy, surgery, age, and tumor size, illuminating their pivotal roles in recurrence likelihood. The second study delved into machine learning techniques, highlighting the prominence of the C5.0 DT algorithm in accurately forecasting breast cancer recurrence. This comparative analysis emphasizes the versatility of machine learning approaches in utilizing clinical data for informed predictions. In the malaria studies, machine learning models effectively

estimated the impact of climatic variables on predicting outbreaks and incidence in The Gambia.

A critical facet of our research has been addressing the challenge of missing data in high-dimensional mixed-type datasets. To this end, we proposed a novel approach integrating Factor Analysis for Mixed Data (FAMD) and k-Nearest Neighbors (kNN) imputation. This technique has shown promising results in imputing missing values in complex datasets, providing a reliable framework for accurate analyses.

8.2 Future Directions

While these studies and the missing data imputation method mark significant advancements, several promising avenues for future research emerge:

- **Integration of Multi-Omics Data:** Incorporating diverse datasets, including genomics, transcriptomics, and proteomics, can provide a more comprehensive understanding of the molecular underpinnings of breast cancer recurrence. Integrating multi-omics data with clinical variables holds immense potential in refining predictive models.
- **Longitudinal Studies and Temporal Patterns:** Extending the temporal horizon of recurrence prediction beyond the three-year window can uncover evolving patterns and dynamic risk factors. Longitudinal studies offer insights into the changing landscape of recurrence risk over time.
- **Diverse Demographic Cohorts:** Generalizing predictive models across diverse

demographic cohorts and geographic regions is crucial for ensuring broad applicability. Studies should aim to encompass populations with varying genetic backgrounds, lifestyles, and healthcare access.

- **Incorporation of Novel Biomarkers:** Exploring emerging biomarkers, such as liquid biopsies and immune profiling, can enhance the precision and sensitivity of predictive models. These biomarkers offer dynamic insights into the evolving tumor microenvironment.
- **Validation in Real-Time Clinical Settings:** Rigorous validation in diverse healthcare settings for both breast cancer and malaria models is imperative for real-world applicability.
- **Ethical and Legal Implications:** As predictive models gain traction in clinical practice, it is paramount to address ethical considerations, including patient consent, privacy, and the responsible use of predictive technologies.
- **Integration with Clinical Decision Support Systems (CDSS):** Integrating predictive models with CDSS can empower healthcare providers with real-time, data-driven insights, facilitating personalized treatment plans and informed patient discussions.
- **Hybridized Ensemble Approach:** Integrating multiple machine learning models into an ensemble approach can potentially yield even more reliable predictions. By leveraging the strengths of different models, this hybridized approach may enhance overall performance.

Incorporation of Additional Predictors: While climatic variables play a crucial role, incorporating other relevant predictors such as treated mosquito nets, indoor residual spray coverage, and socioeconomic factors can provide a more comprehensive understanding of malaria dynamics.

Fine-Grained Spatial and Temporal Analysis: Future studies should focus on predicting malaria cases for specific months and districts, considering the high variability in transmission patterns. This fine-grained analysis will enable targeted interventions in areas with elevated risk.

Integration with Health Information Systems: Linking predictive models with existing health information systems can facilitate real-time monitoring and response. This integration allows for immediate action in the event of predicted outbreaks or surges in malaria cases.

Long-Term Trend Analysis: Examining long-term trends in malaria incidence and outbreaks can offer valuable insights into the effectiveness of control measures. Understanding how these patterns evolve over time is essential for sustained progress.

- Community Engagement and Education: Complementing predictive modeling efforts with community engagement and education initiatives can empower local populations to take proactive measures against malaria. This includes promoting the use of preventive tools and practices.
- Missing Data Imputation Extension: The proposed missing data imputation

method by first applying FAMD before imputation shows promise in enhancing the reliability of our predictive models. Further exploration and validation of this technique across different datasets and domains would be beneficial in establishing its broader applicability.

In conclusion, the collective body of work represents a pivotal step towards personalized healthcare. By embracing the multidimensionality of diseases and utilizing cutting-edge technologies, future endeavors hold the promise of further revolutionizing clinical care and augmenting patient outcomes in the battle against breast cancer and malaria.

REFERENCES

- [1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, “Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA: a cancer journal for clinicians*, vol. 68, no. 6, pp. 394–424, 2018.
- [2] R. B. Nyasa, E. L. Fotabe, and R. N. Ndip, “Trends in malaria prevalence and risk factors associated with the disease in nkongho-mbeng; a typical rural setting in the equatorial rainforest of the south west region of cameroon,” *Plos one*, vol. 16, no. 5, p. e0251380, 2021.
- [3] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, “Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA: a cancer journal for clinicians*, vol. 71, no. 3, pp. 209–249, 2021.
- [4] M. Arnold, E. Morgan, H. Rungay, A. Mafra, D. Singh, M. Laversanne, J. Vignat, J. R. Gralow, F. Cardoso, S. Siesling *et al.*, “Current and future burden of breast cancer: Global statistics for 2020 and 2040,” *The Breast*, vol. 66, pp. 15–23, 2022.

- [5] K. Ravichandran, N. A. Hamdan, and A. Dyab, "Population based survival of female breast cancer cases in riyadh region, saudi arabia," *Asian Pac J Cancer Prev*, vol. 6, no. 1, pp. 72–6, 2005.
- [6] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," *Artificial intelligence in medicine*, vol. 34, no. 2, pp. 113–127, 2005.
- [7] K. D. Miller, L. Nogueira, A. B. Mariotto, J. H. Rowland, K. R. Yabroff, C. M. Alfano, A. Jemal, J. L. Kramer, and R. L. Siegel, "Cancer treatment and survivorship statistics, 2019," *CA: a cancer journal for clinicians*, vol. 69, no. 5, pp. 363–385, 2019.
- [8] R. Peto, J. Boreham, M. Clarke, C. Davies, and V. Beral, "Uk and usa breast cancer deaths down 25% in year 2000 at ages 20-69 years," *Lancet*, vol. 355, no. 9217, p. 1822, 2000.
- [9] H. Pan, R. Gray, J. Braybrooke, C. Davies, C. Taylor, P. McGale, R. Peto, K. I. Pritchard, J. Bergh, M. Dowsett *et al.*, "20-year risks of breast-cancer recurrence after stopping endocrine therapy at 5 years," *New England Journal of Medicine*, vol. 377, no. 19, pp. 1836–1846, 2017.
- [10] S. J. Mandrekar and D. J. Sargent, "Clinical trial designs for predictive biomarker validation: one size does not fit all," *Journal of biopharmaceutical statistics*, vol. 19, no. 3, pp. 530–542, 2009.

- [11] T. Ades, R. Alteri, P. Priti Bandi *et al.*, “American cancer society: Cancer facts and figures,” *Atlanta, GA: American Cancer Society*, 2012.
- [12] R. A. Al-Naggar, “Principles and practice of cancer prevention and control,” *OMICS International: USA*, 2014.
- [13] E. M. Ward, C. E. DeSantis, C. C. Lin, J. L. Kramer, A. Jemal, B. Kohler, O. W. Brawley, and T. Gansler, “Cancer statistics: breast cancer in situ,” *CA: a cancer journal for clinicians*, vol. 65, no. 6, pp. 481–495, 2015.
- [14] C. M. Perou and A.-L. Børresen-Dale, “Systems biology and genomics of breast cancer,” *Cold Spring Harbor perspectives in biology*, vol. 3, no. 2, p. a003293, 2011.
- [15] K. D. Voduc, M. C. Cheang, S. Tyldesley, K. Gelmon, T. O. Nielsen, and H. Kennecke, “Breast cancer subtypes and the risk of local and regional relapse,” *Journal of clinical oncology*, vol. 28, no. 10, pp. 1684–1691, 2010.
- [16] M. C. van Maaren, L. de Munck, L. J. Strobbe, G. S. Sonke, P. J. Westenend, M. L. Smidt, P. M. Poortmans, and S. Siesling, “Ten-year recurrence rates for breast cancer subtypes in the netherlands: a large population-based study,” *International journal of cancer*, vol. 144, no. 2, pp. 263–272, 2019.
- [17] F.-F. Liu, W. Shi, S. J. Done, N. Miller, M. Pintilie, D. Voduc, T. O. Nielsen, S. Nofech-Mozes, M. C. Chang, T. J. Whelan *et al.*, “Identification of a low-risk luminal a breast cancer cohort that may not benefit from breast radiotherapy,” *J Clin Oncol*, vol. 33, no. 18, pp. 2035–2040, 2015.

- [18] C. Criscitiello, H. Azim Jr, P. Schouten, S. Linn, and C. Sotiriou, “Understanding the biology of triple-negative breast cancer,” *Annals of oncology*, vol. 23, pp. vi13–vi18, 2012.
- [19] C. Liedtke, C. Mazouni, K. R. Hess, F. André, A. Tordai, J. A. Mejia, W. F. Symmans, A. M. Gonzalez-Angulo, B. Hennessy, M. Green *et al.*, “Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer,” *Journal of clinical oncology*, vol. 26, no. 8, pp. 1275–1281, 2008.
- [20] R. Dent, M. Trudeau, K. I. Pritchard, W. M. Hanna, H. K. Kahn, C. A. Sawka, L. A. Lickley, E. Rawlinson, P. Sun, and S. A. Narod, “Triple-negative breast cancer: clinical features and patterns of recurrence,” *Clinical cancer research*, vol. 13, no. 15, pp. 4429–4434, 2007.
- [21] M. H. Greene, “Genetics of breast cancer,” in *Mayo Clinic Proceedings*, vol. 72, no. 1. Elsevier, 1997, pp. 54–65.
- [22] Q. Liu, W. Jing, L. Kang, J. Liu, and M. Liu, “Trends of the global, regional and national incidence of malaria in 204 countries from 1990 to 2019 and implications for malaria prevention,” *Journal of Travel Medicine*, vol. 28, no. 5, p. taab046, 2021.
- [23] A. O. Adeyemo, A. T. Aborode, M. A. Bello, A. F. Obianuju, M. M. Hasan, D. O. Kehinde, M. S. Hossain, M. Bardhan, J. O. Imisioluwa, and A. A. Akintola, “Malaria vaccine: The lasting solution to malaria burden in africa,” *Annals of Medicine and Surgery*, vol. 79, p. 104031, 2022.

- [24] A. L. Buczak, B. Baugher, E. Guven, L. C. Ramac-Thomas, Y. Elbert, S. M. Babin, and S. H. Lewis, “Fuzzy association rule mining and classification for the prediction of malaria in south korea,” *BMC medical informatics and decision making*, vol. 15, pp. 1–17, 2015.
- [25] A. Monroe, N. A. Williams, S. Ogoma, C. Karema, and F. Okumu, “Reflections on the 2021 world malaria report and the future of malaria control,” 2022.
- [26] M. Hamid-Adiamoh, D. Nwakanma, B. S. Assogba, M. O. Ndiath, U. D’Alessandro, Y. A. Afrane, and A. Amambua-Ngwa, “Influence of insecticide resistance on the biting and resting preferences of malaria vectors in the gambia,” *Plos one*, vol. 16, no. 6, p. e0241023, 2021.
- [27] E. Jawo, P. A. Mendy, E. Mendy, M. Kah, B. Sanyang *et al.*, “Knowledge, beliefs and attitude towards malaria control and prevention among students in tertiary institutions in the gambia,” *African Journal of Health Sciences*, vol. 35, no. 1, pp. 18–30, 2022.
- [28] J. Mwesigwa, J. Achan, M. Affara, M. Wathuo, A. Worwui, N. I. Mohammed, F. Kanuteh, A. Prom, S. Dierickx, G. L. Di Tanna *et al.*, “Mass drug administration with dihydroartemisinin-piperaquine and malaria transmission dynamics in the gambia: a prospective cohort study,” *Clinical Infectious Diseases*, vol. 69, no. 2, pp. 278–286, 2019.
- [29] H. Broekhuizen, A. Fehr, C. Nieto-Sanchez, J. Muela, K. Peeters-Grietens, T. Smekens, M. Kalleh, E. Rijndertse, J. Achan, and U. D’Alessandro, “Costs

- and barriers faced by households seeking malaria treatment in the upper river region, the gambia,” *Malaria Journal*, vol. 20, no. 1, pp. 1–8, 2021.
- [30] J. Mwesigwa, J. Okebe, M. Affara, G. L. Di Tanna, D. Nwakanma, O. Janha, K. Opondo, K. P. Grietens, J. Achan, and U. D’Alessandro, “On-going malaria transmission in the gambia despite high coverage of control interventions: a nationwide cross-sectional survey,” *Malaria journal*, vol. 14, no. 1, pp. 1–9, 2015.
- [31] M. Mohammadkhani, N. Khanjani, B. Bakhtiari, S. M. Tabatabai, and K. Sheikhzadeh, “The relation between climatic factors and malaria incidence in sisthan and baluchestan, iran,” *Sage Open*, vol. 9, no. 3, p. 2158244019864205, 2019.
- [32] M. C. Thomson, S. J. Mason, T. Phindela, and S. J. Connor, “Use of rainfall and sea surface temperature monitoring for malaria early warning in botswana,” 2005.
- [33] G. Kalipe, V. Gautham, and R. K. Behera, “Predicting malarial outbreak using machine learning and deep learning approach: a review and analysis,” in *2018 International Conference on Information Technology (ICIT)*. IEEE, 2018, pp. 33–38.
- [34] M. Masinde, “Africa’s malaria epidemic predictor: Application of machine learning on malaria incidence and climate data,” in *Proceedings of the 2020 the 4th International Conference on Compute and Data Analysis*, 2020, pp. 29–37.

- [35] T. Santosh, D. Ramesh, and D. Reddy, “Lstm based prediction of malaria abundances using big data,” *Computers in Biology and Medicine*, vol. 124, p. 103859, 2020.
- [36] Y. A. Adamu *et al.*, “Malaria prediction model using machine learning algorithms,” *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 10, pp. 7488–7496, 2021.
- [37] K. Zinszer, R. Kigozi, K. Charland, G. Dorsey, T. F. Brewer, J. S. Brownstein, M. R. Kamya, and D. L. Buckeridge, “Forecasting malaria in a highly endemic country using environmental and clinical predictors,” *Malaria journal*, vol. 14, pp. 1–9, 2015.
- [38] A. B. Pedersen, E. M. Mikkelsen, D. Cronin-Fenton, N. R. Kristensen, T. M. Pham, L. Pedersen, and I. Petersen, “Missing data and multiple imputation in clinical epidemiological research,” *Clinical epidemiology*, pp. 157–166, 2017.
- [39] S. Faisal and G. Tutz, “Nearest neighbor imputation for categorical data by weighting of attributes,” *Information Sciences*, vol. 592, pp. 306–319, 2022.
- [40] N. Z. Abidin and A. R. Ismail, “An improved k-nearest neighbour with grasshopper optimization algorithm for imputation of missing data,” *International Journal of Advances in Intelligent Informatics*, vol. 7, no. 3, pp. 304–317, 2021.
- [41] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2019, vol. 793.

- [42] C. Mack, Z. Su, and D. Westreich, “Managing missing data in patient registries: addendum to registries for evaluating patient outcomes: a user’s guide,” 2018.
- [43] J. W. Graham, “Missing data analysis: Making it work in the real world,” *Annual review of psychology*, vol. 60, pp. 549–576, 2009.
- [44] G. Tutz and S. Ramzan, “Improved methods for the imputation of missing data by nearest neighbor methods,” *Computational Statistics & Data Analysis*, vol. 90, pp. 84–99, 2015.
- [45] S. Faisal and G. Tutz, “Imputation methods for high-dimensional mixed-type datasets by nearest neighbors,” *Computers in Biology and Medicine*, vol. 135, p. 104577, 2021.
- [46] T. Mitchel, “Machine learning, mcgraw-hill education (ise editions),” 1997.
- [47] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, “Machine learning applications in cancer prognosis and prediction,” *Computational and structural biotechnology journal*, vol. 13, pp. 8–17, 2015.
- [48] S. Gupta, D. Kumar, and A. Sharma, “Data mining classification techniques applied for breast cancer diagnosis and prognosis,” *Indian Journal of Computer Science and Engineering (IJCSE)*, vol. 2, no. 2, pp. 188–195, 2011.
- [49] A. Nair, A. Gupta, M. Dalal, and S. Levine, “Awac: Accelerating online reinforcement learning with offline datasets,” *arXiv preprint arXiv:2006.09359*, 2020.

- [50] G. Meyfroidt, F. Güiza, J. Ramon, and M. Bruynooghe, “Machine learning techniques to examine large patient databases,” *Best Practice & Research Clinical Anaesthesiology*, vol. 23, no. 1, pp. 127–143, 2009.
- [51] A. E. Mohamed, “Comparative study of four supervised machine learning techniques for classification,” *International Journal of Applied*, vol. 7, no. 2, pp. 1–15, 2017.
- [52] B. Gupta, A. Rawat, A. Jain, A. Arora, and N. Dhimi, “Analysis of various decision tree algorithms for classification in data mining,” *International Journal of Computer Applications*, vol. 163, no. 8, pp. 15–19, 2017.
- [53] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013, vol. 112.
- [54] L. Breiman, J. Friedman, R. Olshen, and C. Stone, “Classification and regression trees. monterey, ca: Wadsworth & brooks,” 1984.
- [55] G. V. Kass, “An exploratory technique for investigating large quantities of categorical data,” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 29, no. 2, pp. 119–127, 1980.
- [56] W.-Y. Loh and Y.-S. Shih, “Split selection methods for classification trees,” *Statistica sinica*, pp. 815–840, 1997.
- [57] J. R. Quinlan, “Induction of decision trees,” *Machine learning*, vol. 1, pp. 81–106, 1986.

- [58] ———, “Program for machine learning,” *C4*. 5, 1993.
- [59] L. Breiman, “Random forests,” *Machine learning*, vol. 45, pp. 5–32, 2001.
- [60] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967.
- [61] C. Eyupoglu, “Breast cancer classification using k-nearest neighbors algorithm,” *The Online Journal of Science and Technology*, vol. 8, no. 3, pp. 29–34, 2018.
- [62] R. Rawal, “Breast cancer prediction using machine learning,” *Journal of Emerging Technologies and Innovative Research (JETIR)*, vol. 13, no. 24, p. 7, 2020.
- [63] S. Zhang, X. Li, M. Zong, X. Zhu, and D. Cheng, “Learning k for knn classification,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 8, no. 3, pp. 1–19, 2017.
- [64] V. Ramalingam, A. Dandapath, and M. K. Raja, “Heart disease prediction using machine learning techniques: a survey,” *International Journal of Engineering & Technology*, vol. 7, no. 2.8, pp. 684–687, 2018.
- [65] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 1999.
- [66] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene selection for cancer classification using support vector machines,” *Machine learning*, vol. 46, pp. 389–422, 2002.

- [67] J. Fan, X. Wang, L. Wu, H. Zhou, F. Zhang, X. Yu, X. Lu, and Y. Xiang, "Comparison of support vector machine and extreme gradient boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in china," *Energy conversion and management*, vol. 164, pp. 102–111, 2018.
- [68] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC medical informatics and decision making*, vol. 19, no. 1, pp. 1–16, 2019.
- [69] A. Tharwat, "Parameter investigation of support vector machine classifier with kernel functions," *Knowledge and Information Systems*, vol. 61, pp. 1269–1302, 2019.
- [70] C. Campbell, "Kernel methods: a survey of current techniques," *Neurocomputing*, vol. 48, no. 1-4, pp. 63–84, 2002.
- [71] M. van Gerven and S. M. Bohte, "Artificial neural networks as models of neural information processing," 2017.
- [72] S. Kharya, D. Dubey, and S. Soni, "Predictive machine learning techniques for breast cancer detection," *International journal of computer science and information Technologies*, vol. 4, no. 6, pp. 1023–1028, 2013.
- [73] S. B. Maind, P. Wankar *et al.*, "Research paper on basic of artificial neural network," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 2, no. 1, pp. 96–100, 2014.

- [74] R. Dastres and M. Soori, “Artificial neural network systems,” *International Journal of Imaging and Robotics (IJIR)*, vol. 21, no. 2, pp. 13–25, 2021.
- [75] A. K. Dwivedi, “Performance evaluation of different machine learning techniques for prediction of heart disease,” *Neural Computing and Applications*, vol. 29, pp. 685–693, 2018.
- [76] R. M. Sadek, S. A. Mohammed, A. R. K. Abunbehan, A. K. H. A. Ghattas, M. R. Badawi, M. N. Mortaja, B. S. Abu-Nasser, and S. S. Abu-Naser, “Parkinson’s disease prediction using artificial neural network,” 2019.
- [77] M. D. Ganggayah, N. A. Taib, Y. C. Har, P. Lio, and S. K. Dhillon, “Predicting factors for survival of breast cancer patients using machine learning techniques,” *BMC medical informatics and decision making*, vol. 19, pp. 1–17, 2019.
- [78] T. Chen and C. Guestrin, “Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining,” 2016.
- [79] —, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.
- [80] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, “A comparative analysis of gradient boosting algorithms,” *Artificial Intelligence Review*, vol. 54, pp. 1937–1967, 2021.

- [81] I. Tougui, A. Jilbab, and J. El Mhamdi, “Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications,” *Healthcare informatics research*, vol. 27, no. 3, pp. 189–199, 2021.
- [82] S. Arlot and A. Celisse, “A survey of cross-validation procedures for model selection,” 2010.
- [83] Y. Jiao and P. Du, “Performance measures in evaluating machine learning based bioinformatics predictors for classifications,” *Quantitative Biology*, vol. 4, pp. 320–330, 2016.
- [84] O. Theobald, *Machine learning for absolute beginners: a plain English introduction*. Scatterplot press London, UK, 2017, vol. 157.
- [85] M. Hossin and M. N. Sulaiman, “A review on evaluation metrics for data classification evaluations,” *International journal of data mining & knowledge management process*, vol. 5, no. 2, p. 1, 2015.
- [86] W. Kim, K. S. Kim, J. E. Lee, D.-Y. Noh, S.-W. Kim, Y. S. Jung, M. Y. Park, and R. W. Park, “Development of novel breast cancer recurrence prediction model using support vector machine,” *Journal of breast cancer*, vol. 15, no. 2, pp. 230–238, 2012.
- [87] C. Boeri, C. Chiappa, F. Galli, V. De Berardinis, L. Bardelli, G. Carcano, and F. Rovera, “Machine learning techniques in breast cancer prognosis prediction: A primary evaluation,” *Cancer medicine*, vol. 9, no. 9, pp. 3234–3243, 2020.

- [88] L. G. Ahmad, A. Eshlaghy, A. Poorebrahimi, M. Ebrahimi, A. Razavi *et al.*, “Using three machine learning techniques for predicting breast cancer recurrence,” *J Health Med Inform*, vol. 4, no. 124, p. 3, 2013.
- [89] D. Umesh and B. Ramachandra, “Association rule mining based predicting breast cancer recurrence on seer breast cancer data,” in *2015 International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT)*. IEEE, 2015, pp. 376–380.
- [90] Q. Fan, C.-j. Zhu, and L. Yin, “Predicting breast cancer recurrence using data mining techniques,” in *2010 International Conference on Bioinformatics and Biomedical Technology*. IEEE, 2010, pp. 310–311.
- [91] T. Al-Quraishi, J. H. Abawajy, M. U. Chowdhury, S. Rajasegarar, and A. S. Abdalrada, “Breast cancer recurrence prediction using random forest model,” in *Recent Advances on Soft Computing and Data Mining: Proceedings of the Third International Conference on Soft Computing and Data Mining (SCDM 2018), Johor, Malaysia, February 06-07, 2018*. Springer, 2018, pp. 318–329.
- [92] A. I. Pritom, M. A. R. Munshi, S. A. Sabab, and S. Shihab, “Predicting breast cancer recurrence using effective classification and feature selection technique,” in *2016 19th International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2016, pp. 310–314.
- [93] C.-C. Chang and S.-H. Chen, “Developing a novel machine learning-based classification scheme for predicting spcs in breast cancer survivors,” *Frontiers in*

Genetics, vol. 10, p. 848, 2019.

- [94] L. Macías-García, M. Martinez-Ballesteros, J. M. Luna-Romera, J. M. García-Heredia, J. García-Gutiérrez, and J. C. Riquelme-Santos, “Autoencoded dna methylation data to predict breast cancer recurrence: Machine learning models and gene-weight significance,” *Artificial Intelligence in Medicine*, vol. 110, p. 101976, 2020.
- [95] A. I. Aishwarja, N. J. Eva, S. Mushtary, Z. Tasnim, N. I. Khan, and M. N. Islam, “Exploring the machine learning algorithms to find the best features for predicting the breast cancer and its recurrence,” in *Intelligent Computing and Optimization: Proceedings of the 3rd International Conference on Intelligent Computing and Optimization 2020 (ICO 2020)*. Springer, 2021, pp. 546–558.
- [96] O. Nkiruka, R. Prasad, and O. Clement, “Prediction of malaria incidence using climate variability and machine learning,” *Informatics in Medicine Unlocked*, vol. 22, p. 100508, 2021.
- [97] Y. W. Lee, J. W. Choi, and E.-H. Shin, “Machine learning model for predicting malaria using clinical information,” *Computers in Biology and Medicine*, vol. 129, p. 104151, 2021.
- [98] A. Stephen, P. O. Akomolafe, and K. I. Ogundoyin, “A model for predicting malaria outbreak using machine learning technique,” *Annals. Computer Science Series*, vol. 19, no. 1, 2021.

- [99] S. R. Christophers, “Epidemic malaria of the punjab: with a note of a method of predicting epidemic years,” *Trans Committee Stud Malaria India*, vol. 2, pp. 17–26, 1911.
- [100] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, “Missing value estimation methods for dna microarrays,” *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.
- [101] D. J. Stekhoven and P. Bühlmann, “Missforest—non-parametric missing value imputation for mixed-type data,” *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2012.
- [102] Y. Zhao and Q. Long, “Multiple imputation in the presence of high-dimensional data,” *Statistical Methods in Medical Research*, vol. 25, no. 5, pp. 2021–2035, 2016.
- [103] A. D. Shah, J. W. Bartlett, J. Carpenter, O. Nicholas, and H. Hemingway, “Comparison of random forest and parametric imputation models for imputing missing data using mice: a caliber study,” *American journal of epidemiology*, vol. 179, no. 6, pp. 764–774, 2014.
- [104] S. G. Liao, Y. Lin, D. D. Kang, D. Chandra, J. Bon, N. Kaminski, F. C. Sciurba, and G. C. Tseng, “Missing value imputation in high-dimensional phenomic data: imputable or not, and how?” *BMC bioinformatics*, vol. 15, no. 1, pp. 1–12, 2014.
- [105] A. De La Rochefordiere, F. Campana, J. Fenton, J. Vilcoq, A. Fourquet, B. Asselain, S. Scholl, P. Pouillart, J.-C. Durand, and H. Magdelenat, “Age as prog-

- nostic factor in premenopausal breast carcinoma,” *The Lancet*, vol. 341, no. 8852, pp. 1039–1043, 1993.
- [106] P. P. Rosen, M. L. Lesser, D. W. Kinne, and E. J. Beattie, “Breast carcinoma in women 35 years of age or younger,” *Annals of surgery*, vol. 199, no. 2, p. 133, 1984.
- [107] R. Yancik, M. N. Wesley, L. A. Ries, R. J. Havlik, B. K. Edwards, and J. W. Yates, “Effect of age and comorbidity in postmenopausal breast cancer patients aged 55 years and older,” *Jama*, vol. 285, no. 7, pp. 885–892, 2001.
- [108] B. C. Yankaskas, “Epidemiology of breast cancer in young women,” *Breast disease*, vol. 23, no. 1, pp. 3–8, 2006.
- [109] A. Barchielli and D. Balzi, “Age at diagnosis, extent of disease and breast cancer survival: a population-based study in florence, italy,” *Tumori Journal*, vol. 86, no. 2, pp. 119–123, 2000.
- [110] V. T. DeVita, T. S. Lawrence, and S. A. Rosenberg, *DeVita, Hellman, and Rosenberg’s cancer: principles & practice of oncology*. Lippincott Williams & Wilkins, 2008, vol. 2.
- [111] U. Veronesi, G. Paganelli, G. Viale, A. Luini, S. Zurrada, V. Galimberti, M. Intra, P. Veronesi, C. Robertson, P. Maisonneuve *et al.*, “A randomized comparison of sentinel-node biopsy with routine axillary dissection in breast cancer,” *New England Journal of Medicine*, vol. 349, no. 6, pp. 546–553, 2003.

- [112] E. B. C. T. C. Group *et al.*, “Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials,” *The Lancet*, vol. 365, no. 9472, pp. 1687–1717, 2005.
- [113] E. Tsangaris, A. F. Klassen, M. N. Kaur, S. Voineskos, L. Bordeleau, T. Zhong, J. Broyles, and A. L. Pusic, “Development and psychometric validation of the breast-q sensation module for women undergoing post-mastectomy breast reconstruction,” *Annals of Surgical Oncology*, vol. 28, no. 12, pp. 7842–7853, 2021.
- [114] S. M. Swain, S.-B. Kim, J. Cortés, J. Ro, V. Semiglazov, M. Campone, E. Ciruelos, J.-M. Ferrero, A. Schneeweiss, A. Knott *et al.*, “Pertuzumab, trastuzumab, and docetaxel for her2-positive metastatic breast cancer (cleopatra study): overall survival results from a randomised, double-blind, placebo-controlled, phase 3 study,” *The lancet oncology*, vol. 14, no. 6, pp. 461–471, 2013.
- [115] C. K. Anders, D. S. Hsu, G. Broadwater, C. R. Acharya, J. A. Foekens, Y. Zhang, Y. Wang, P. K. Marcom, J. R. Marks, P. G. Febbo *et al.*, “Young age at diagnosis correlates with worse prognosis and defines a subset of breast cancers with shared patterns of gene expression,” *Journal of clinical oncology*, vol. 26, no. 20, pp. 3324–3330, 2008.
- [116] A. H. Partridge, M. E. Hughes, E. T. Warner, R. A. Ottesen, Y.-N. Wong, S. B. Edge, R. L. Theriault, D. W. Blayney, J. C. Niland, E. P. Winer *et al.*, “Subtype-dependent relationship between young age at diagnosis and breast

- cancer survival,” *Journal of Clinical Oncology*, vol. 34, no. 27, pp. 3308–3314, 2016.
- [117] B. Weigelt, A. Mackay, R. A’hern, R. Natrajan, D. S. Tan, M. Dowsett, A. Ashworth, and J. S. Reis-Filho, “Breast cancer molecular profiling with single sample predictors: a retrospective analysis,” *The lancet oncology*, vol. 11, no. 4, pp. 339–349, 2010.
- [118] L. González-Castro, M. Chávez, P. Duflo, V. Bleret, A. G. Martin, M. Zobel, J. Nateqi, S. Lin, J. J. Pazos-Arias, G. Del Fiol *et al.*, “Machine learning algorithms to predict breast cancer recurrence using structured and unstructured sources from electronic health records,” *Cancers*, vol. 15, no. 10, p. 2741, 2023.
- [119] K. Park, A. Ali, D. Kim, Y. An, M. Kim, and H. Shin, “Robust predictive model for evaluating breast cancer survivability,” *Engineering Applications of Artificial Intelligence*, vol. 26, no. 9, pp. 2194–2205, 2013.
- [120] H. Wickham and G. Grolemund, *R for data science: import, tidy, transform, visualize, and model data.* ” O’Reilly Media, Inc.”, 2016.
- [121] M. K. Goel, P. Khanna, and J. Kishore, “Understanding survival analysis: Kaplan-meier estimate,” *International journal of Ayurveda research*, vol. 1, no. 4, p. 274, 2010.
- [122] C. DR, “Oakes d. analysis of survival data,” *London, Chapman & Hill*, 1984.

- [123] M. Bacauskiene and A. Verikas, “Selecting salient features for classification based on neural network committees,” *Pattern recognition letters*, vol. 25, no. 16, pp. 1879–1891, 2004.
- [124] S. B. Sakri, N. B. A. Rashid, and Z. M. Zain, “Particle swarm optimization feature selection for breast cancer recurrence prediction,” *IEEE Access*, vol. 6, pp. 29 637–29 647, 2018.
- [125] S. O. Bittaye, A. Jagne, L. E. Jaiteh, B. Nadjm, A. Amambua-Ngwa, A. K. Sesay, Y. Singhateh, E. Effa, O. Nyan, and R. Njie, “Clinical manifestations and outcomes of severe malaria in adult patients admitted to a tertiary hospital in the gambia,” *Malaria journal*, vol. 21, no. 1, pp. 1–8, 2022.
- [126] K. A. Collins, S. Ceesay, S. Drammeh, F. K. Jaiteh, M. A. Guery, K. Lanke, L. Grignard, W. Stone, D. J. Conway, U. D’alessandro *et al.*, “A cohort study on the duration of plasmodium falciparum infections during the dry season in the gambia,” *The Journal of infectious diseases*, vol. 226, no. 1, pp. 128–137, 2022.
- [127] J. Mwesigwa, J. Achan, G. L. Di Tanna, M. Affara, M. Jawara, A. Worwui, M. Hamid-Adiamoh, F. Kanuteh, S. Ceesay, T. Bousema *et al.*, “Residual malaria transmission dynamics varies across the gambia despite high coverage of control interventions,” *PLoS One*, vol. 12, no. 11, p. e0187059, 2017.
- [128] D. Singh and B. Singh, “Investigating the impact of data normalization on classification performance,” *Applied Soft Computing*, vol. 97, p. 105524, 2020.

- [129] N. Meinshausen and G. Ridgeway, “Quantile regression forests.” *Journal of machine learning research*, vol. 7, no. 6, 2006.
- [130] N. A. B. Klutse, F. Aboagye-Antwi, K. Owusu, Y. Ntiamoa-Baidu *et al.*, “Assessment of patterns of climate variables and malaria cases in two ecological zones of ghana,” *Open Journal of Ecology*, vol. 4, no. 12, p. 764, 2014.
- [131] L. A. Moreira, J. Wang, F. H. Collins, and M. Jacobs-Lorena, “Fitness of anopheline mosquitoes expressing transgenes that inhibit plasmodium development,” *Genetics*, vol. 166, no. 3, pp. 1337–1341, 2004.
- [132] K. Ousman, A. Jimoh O., and M. P. Hossain, “Predicting malaria outbreak in the gambia using machine learning techniques,” *working paper*, 2023.
- [133] O. Ebhuoma, M. Gebreslasie, and L. Magubane, “A seasonal autoregressive integrated moving average (sarima) forecasting model to predict monthly malaria cases in kwazulu-natal, south africa,” *South African medical journal*, vol. 108, no. 7, 2018.
- [134] J. Zhai, Q. Lu, W. Hu, S. Tong, B. Wang, F. Yang, Z. Xu, S. Xun, and X. Shen, “Development of an empirical model to predict malaria outbreaks based on monthly case reports and climate variables in hefei, china, 1990–2011,” *Acta tropica*, vol. 178, pp. 148–154, 2018.
- [135] I. T. Jolliffe and J. Cadima, “Principal component analysis: a review and recent developments,” *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, 2016.

- [136] S. Van Buuren and K. Groothuis-Oudshoorn, “mice: Multivariate imputation by chained equations in r,” *Journal of statistical software*, vol. 45, pp. 1–67, 2011.

VITAE

Address: MDI Road, Kanifing, P.O Box 35350, Serrekunda, The Gambia

Phone: +220 3512066

Email: ouskhan95@gmail.com

Objective

Statistician with a passion for learning, developing skills in statistics, and contributing to collaborative environments.

Skills

- Competent in R programming, MINITAB, and MATLAB
- Proficient in Microsoft Office (Excel, Word, PowerPoint)
- Strong love for teaching, classroom teaching, and data interpretation

Work History

Recitation Teaching Assistant

King Fahd University of Petroleum and Minerals, Dammam

September 2021 - Current

- Grade quizzes, homework, and exams designated by my instructor
- Take attendance in every class
- Conduct problem-solving class for each section weekly
- Proctor exams

Census Field Supervisor

Gambia Bureau of Statistics, Banjul

November 2020 - February 2021

- Distributed assignments to each enumerator
- Reviewed data collected by enumerators
- Provided on-the-job training and other assistance for enumerators
- Checked on enumerator's while conducting interviews to check if they were asking the right questions
- Sent in completed and reviewed questionnaires to head office daily

Enumerator

Gambia Bureau of Statistics, Banjul

October 2019 - March 2020

- Traveled door-to-door in assigned communities and requested participation in the survey interview, clearly explaining process and purpose
- Recognized and gave an account of problems in obtaining data
- Reviewed questionnaire submissions for accuracy and completeness at end of each day, flagging incomplete or problematic surveys for further review
- Executed follow-up visits to houses to resolve survey issues or obtain additional information, expressing appreciation for participation and regret for inconvenience

Auditor

Zenith Bank, Banjul

February 2019 - October 2019

- Performed objective, independent, and reliable assessments of the effectiveness of the bank's risk management activities
- Responsible for resolving material issues with senior management, reporting audit results, and giving relevant recommendations
- Examined and evaluated the effectiveness and adequacy of the internal control system

- Reviewed financial reports and accounting records for accuracy and reliability

Education

Master of Science: Applied Statistics

King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia

September 2021 - December 2023

Bachelor of Science: Mathematics

University of The Gambia, Kanifing, The Gambia

September 2014 - July 2018

Research Experience

- I.** Worked on a research paper (Prognostic Model for Predicting Recurrence in Breast Cancer Patients in Saudi Arabia) - Submitted in Journal of Clinical Oncology.
- II.** Working on a research paper (Quantile Random Forest-Based Predictive Modeling of Malaria Incidence in The Gambia: Integrating Climatic and Non-Climatic Factors) - In progress for publication.
- III.** Worked on a research paper (Predicting malaria outbreak in The Gambia using machine learning techniques) - Submitted in PLOS One journal.

- IV.** Working on a research paper (Comparative Analysis of different Machine Learning Techniques in Predicting Breast Cancer Recurrence among Patients in the Kingdom of Saudi Arabia) - In progress for publication.

Leadership Experience

President, Kuntaya Ex Students' Association

August 2016 - September 2021

- Presided over all the association matters and meetings
- Coordinated all the functions of the other officers and delegated to them, when necessary, such duties as are germane to the smooth running of the association
- Encouraged mutual understanding among all members

Head boy, Kuntaya Basic Cycle School

May 2010 - May 2011

- Represented the school in events like debate and quiz competition
- Planned and assigned duties to the senior and class prefects, headed meetings, explained the rules and regulations, and was often the first in line for resolving conflicts
- Spoke with teachers and the principal to discuss any issues affecting students

References

Available upon request.