

RISK FACTOR PREDICTION AND SURVIVAL ANALYSIS FOR
CHRONIC DISEASE RECURRENCE USING MACHINE LEARNING
MODELS

BY

Maha Abd El Hakeem Abd El Hameed Said

A Thesis Presented to the
COLLEGE OF GRADUATE AND INTERDISCIPLINARY 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

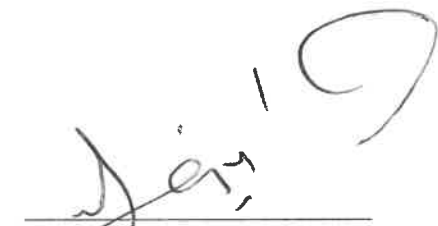
May 2025

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN- 31261, SAUDI ARABIA

DEANSHIP OF GRADUATE STUDIES

This thesis, written by Maha Abd El Hakeem Abd El Hameed Said under the direction of her thesis advisor and approved by her thesis committee, has been presented and accepted by the Dean of Graduate Studies, in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE IN APPLIED STATISTICS.**



Prof. Monther Rashed Saleh Alfuraidan
Department Chairman

Prof. Suliman Al-Homidan
Dean of Graduate Studies

Date



Dr. Ridwan Adeyemi Sanusi
(Advisor)

Prof. Mohammad Hafidz Omar
(Member)

Dr. Hani Almohanna
(Member)

© Maha Abd El Hakeem Abd El Hameed Said

2025

To my loving family, whose support has always been my strength. To my husband— my rock, my anchor, and unwavering support — your belief in me made this achievement possible and gave me the courage to keep going. To my beloved children, Adam, Ahmed, and Aseel — your love, joy, and patience have been my light in the darkest of nights, and my constant source of inspiration. I dedicate this thesis to you all, with deepest gratitude and endless love.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to my supervisor, Dr. Ridwan Adeyemi Sanusi, for his continuous encouragement, insightful feedback, and dedicated supervision, which were instrumental in shaping the direction and quality of this thesis. I am also deeply grateful to my former supervisor, Dr. Jimoh Ajadi, whose early mentorship and steadfast academic support laid the foundation for my scholarly development and whose continued guidance and intellectual contributions significantly enriched this research.

I am sincerely grateful to my thesis committee for their valuable insights and support. I also thank Ms. Norah Aldosary and the hospital team in Dammam for facilitating access to the colorectal cancer dataset, which contributed to the completion of this study.

I would also like to acknowledge the Department of Mathematics and Statistics for offering a supportive academic environment and the scholarship that enabled me to carry out this work.

Lastly, I extend my deepest thanks to my dear parents for their constant prayers, love, and the strong values they instilled in me. To my loving husband, whose unwavering support, patience, and encouragement were my greatest source of strength. His support was vital to the completion of this thesis. He truly deserves my heartfelt appreciation and recognition, along with my everlasting love. And to my beloved children, you are the light of my life and my greatest inspiration. Your smiles, laughter, and unconditional love gave me the strength to keep going.

TABLE OF CONTENTS

ACKNOWLEDGMENT	V
LIST OF TABLES	X
LIST OF FIGURES	XII
ABSTRACT (ENGLISH)	XIII
ملخص الرسالة	XV
1. INTRODUCTION.....	1
1.1 Background.....	1
1.2 Literature review.....	5
1.3 Problem Description	11
1.4 Objective.....	12
1.5 Methodology.....	13
1.6 Unique Contributions and Innovations	17
2. MACHINE LEARNING FRAMEWORK AND METHODOLOGICAL FOUNDATIONS.....	19
2.1 Machine Learning Techniques.....	19
2.1.1 Logistic Regression (LR).....	20
2.1.2 Decision Trees (DT)	22
2.1.3 Random Forest (RF)	25
2.1.4 Support Vector Machines (SVM)	26
2.1.5 Artificial Neural Networks (ANNs).....	29
2.1.6 Extreme Gradient Boosting (XGBoost).....	32
2.1.7 Ensemble Modelling	33

2.2 Machine Learning Evaluation Techniques	35
2.2.1 Hold-Out Method	35
2.2.2 k-fold cross-validation	36
2.2.3 Leave-One-Out Cross-Validation (LOOCV):.....	37
2.2.4 Repeated K-Fold Cross-Validation.....	38
2.2.5 Stratified Cross-Validation	38
2.3 Bias-Variance Trade-off in Model Evaluation	39
2.3.1 Bias-Variance Trade-off and Its Impact on Model Performance.....	40
2.3.2 Bias-Variance Trade-off in Different Machine Learning Models	41
2.3.3 Strategies to Balance Bias and Variance	42
2.4 Evaluation of Machine Learning Performance	42
2.4.1 Classification Model Evaluation Metrics.....	42
2.4.2 Performance Metrics for Regression Models	46
3. STATISTICAL MODELING OF COLORECTAL CANCER OUTCOMES: CANCER STATUS, RECURRENCE, AND VITAL STATUS.....	48
3.1 Background.....	48
3.2 Methods	49
3.2.1 Study Design and Participants	49
3.2.2 Outcome Variables	49
3.2.3 Predictor Variables	50
3.2.4 Inclusion and Exclusion Criteria.....	53
3.2.5 Statistical Analysis	53
3.3 Result.....	57
3.3.1 Descriptive Characteristics of the Cohort.....	57
3.3.2 Multivariate Logistic Regression Analysis for Cancer Status	67
3.3.3 Multivariate Logistic Regression Analysis for Recurrence Status.....	70

3.3.4 Multivariate Logistic Regression Analysis for Vital Status.....	72
3.3.5 Multivariable Cox Proportional Hazards Analysis for cancer status	75
3.3.6 Multivariable Cox Proportional Hazards Analysis for Recurrence	78
3.3.7 Multivariable Cox Proportional Hazards Analysis for Overall Survival (Vital Status).....	80
3.4 Discussion.....	84

4. MACHINE LEARNING MODELS FOR PREDICTING KEY OUTCOMES IN CRC PATIENTS IN SAUDI ARABIA 87

4.1 Background.....	87
4.2 Methods	87
4.2.1 Dataset	88
4.2.2 Feature Selection for Prognostic Modeling	88
4.2.3 Validation Strategy and Performance Evaluation Metrics.....	89
4.3 Result.....	90
4.3.1 Cancer Status Prediction.....	90
4.3.2 Recurrence Prediction.....	95
4.3.3 Vital Status Prediction	99
4.4 Discussion.....	103

5. THE IMPACT OF TRANSITIONS IN CANCER CONTROL POLICIES ON COLORECTAL CANCER PREVALENCE, INCIDENCE, AND MORTALITY.. 106

5.1 Background.....	106
5.2 Methods	108
5.2.1 Data source, exclusion, and inclusion approach	108
5.2.2 Defining Study Periods by Policy Transitions:.....	109
5.2.3 Statistical analyses	112
5.3 Results	115
5.3.1 Characteristics of the Study Population.....	115

5.3.2 Goodness of fit tests for regression models	116
5.3.3 Age- and sex-adjusted prevalence, incidence and mortality	117
5.3.4 Result.....	120
5.4 Discussion.....	123
6. CONCLUSION AND PROSPECTS FOR FUTURE RESEARCH	125
6.1 Conclusion.....	125
6.2 Prospects for Future Research	126
REFERENCES.....	128
VITAE	146

LIST OF TABLES

Table 3.1: Distribution of the Three Primary Outcome Variables.....	50
Table 3.2: Description of the variables in the dataset.	55
Table 3.3: Comparison of patients' characteristics between the No Evidence of Disease (NED) And Evidence of Disease (EOD) cases.....	59
Table 3.4: Comparison of patients' characteristics between the recurrent and non-recurrent cases.....	62
Table 3.5: Comparison of patients' characteristics between two categories of Vital Status "Alive" and "Dead" Cases.....	65
Table 3.6: The result of the multivariate logistic regression model for cancer status with the adjusted odd ratio with 95% CI.....	68
Table 3.7: The result of the multivariate logistic regression model for recurrence with the adjusted odd ratio with 95% CI.....	70
Table 3.8: The result of the multivariate logistic regression model for vital with the adjusted odd ratio with 95% CI.	73
Table 3.9: Results of the Multivariate Cox Proportional Hazards Model for Cancer Status with Adjusted Hazard Ratios with 95% CI.....	76
Table 3.10: Results of the Multivariate Cox Proportional Hazards Model for Recurrence Status with Adjusted Hazard Ratios with 95% CI.	79
Table 3.11: Results of the Multivariate Cox Proportional Hazards Model for Vital Status with Adjusted Hazard Ratios with 95% CI.....	81
Table 4.1: Training Set Evaluation Metrics for Machine Learning Models Predicting Cancer Status	91
Table 4.2: Validation Set Evaluation Metrics for Machine Learning Models Predicting Cancer Status	92
Table 4.3: Features Retained in Final Machine Learning Models for Cancer Status Prediction	94
Table 4.4: Training Set Evaluation Metrics for Machine Learning Models Predicting CRC Recurrence	96
Table 4.5: Validation Set Evaluation Metrics for Machine Learning Models Predicting CRC Recurrence	96
Table 4.6: Features Retained in Final Machine Learning Models for CRC Recurrence Prediction	99
Table 4.7: Training Set Evaluation Metrics for Machine Learning Models Predicting Vital Status in CRC Patients.....	100
Table 4.8: Validation Set Evaluation Metrics for Machine Learning Models Predicting Vital Status in CRC Patients.....	100
Table 4.9: Features Retained in Final Machine Learning Models for Vital Status Prediction	103

Table 5.1: Demographic Characteristics of the Study Population in Selected Years	116
Table 5. 2: Prevalence, incidence and mortality and their average annual rate of change (%).....	119

LIST OF FIGURES

Figure 2.1: The impact of model complexity on bias, variance, and total error.....	41
Figure 3.1: ROC curve of the logistic regression model showing the AUC values of the training and testing set for Cancer Status.....	70
Figure 3.2: ROC curve of the logistic regression model showing the AUC values of the training and testing set for Recurrence Status.....	72
Figure 3.3: ROC curve of the logistic regression model showing the AUC values of the training and testing set for Vital Status.....	75
Figure 3.4: Observed vs. Predicted Kaplan–Meier Curves Cancer Status Survival curve	78
Figure 3.5: Observed vs. Predicted Kaplan–Meier Curves for Recurrence-Free Survival	80
Figure 3.6: Observed vs. Predicted Kaplan–Meier Survival Curves for Overall Survival in CRC Patients.....	84
Figure 4.1: ROC Curves of Machine Learning Models for Predicting Cancer Status in CRC Patients (Training Set)	93
Figure 4.2: ROC Curves of Machine Learning Models for Predicting CRC Recurrence in the Training Set.....	98
Figure 4.3: ROC Curves of Machine Learning Models for Predicting Vital Status in Colorectal Cancer Patients (Training Set)	102
Figure 5.1: Comparing Model Fit for Poisson and Negative Binomial Regression for prevalence, incidence and mortality.	117
Figure 5.2: Control Chart for Prevalence, Incidence and Mortality; Horizontal dashed line represents the upper control limit of 24.2; Vertical dashed lines represent the end of the transition periods.....	122

ABSTRACT (ENGLISH)

Full Name : [Maha Abd El Hakeem Abd El Hameed Said]

Thesis Title : [RISK FACTOR PREDICTION AND SURVIVAL ANALYSIS FOR
CHRONIC DISEASE RECURRENCE USING MACHINE
LEARNING MODELS]

Major Field : Applied Statistics]

Date of Degree : [May, 2025]

Chronic diseases pose a significant public health burden in the Kingdom of Saudi Arabia (KSA), with colorectal cancer (CRC) being the most commonly diagnosed cancer among men and the third among women. This thesis introduces a data-driven approach to CRC outcome prediction by modeling three clinically relevant endpoints: cancer status, recurrence and vital status. Using hospital-based clinical data, this study explores traditional statistical models, namely Cox proportional hazards regression and Logistic Regression, and advanced machine learning techniques, such as support vector machines, random forests, Extreme Gradient Boosting and Artificial Neural Networks, to develop accurate and interpretable predictive models.

Descriptive statistics are used to profile patient characteristics, while model performance is evaluated using cross-validation and metrics including AUC, sensitivity, specificity, and predictive values. Feature selection was performed in both traditional and machine learning models—using stepwise selection in traditional methods and backward elimination within nested cross-validation for machine learning models—to enhance model efficiency and interpretability. Among the machine learning models, XGBoost and ANN showed the most balanced and robust performance in predicting cancer status and survival, while Random Forest and XGBoost demonstrated strong utility for recurrence prediction. Logistic and Cox models confirmed the importance of tumor site, stage, lymph node involvement, and treatment interventions (surgery, chemotherapy, immunotherapy) across all outcomes.

A separate chapter examines the impact of transitions in national cancer control policies on CRC prevalence, incidence, and mortality in KSA over the past three decades. This component uses Global Burden of Disease 2021 data (1990–2021), applying negative

binomial regression and Hotelling's T^2 -based control charts to detect shifts in cancer trends associated with policy changes. Significant trend deviations were detected in 1995 and 2021, corresponding with the launch of the Saudi Cancer Registry and the expansion of CRC screening programs. While screening led to increased detection, mortality trends showed limited improvement, highlighting gaps in treatment effectiveness and the need for strengthened follow-up care.

The study's primary contribution lies in its multi-outcome modeling framework, offering insights that extend beyond single-endpoint prediction. By bridging clinical analytics with public health evaluation, this research aims to enhance risk stratification, support personalized care, and inform national cancer control efforts, ultimately contributing to improved CRC management in KSA.

ملخص الرسالة

الاسم الكامل: مها عبد الحكيم عبد الحميد سعيد

عنوان الرسالة: التنبؤ بعوامل الخطر وتحليل البقاء لتكرار الإصابة بالأمراض المزمنة باستخدام نماذج تعلم الآلة

التخصص: الإحصاء التطبيقي

تاريخ الدرجة العلمية: مايو ٢٠٢٥

تشكل الأمراض المزمنة عبئًا صحيًا عالميًا كبيرًا في المملكة العربية السعودية، ويُعد سرطان القولون والمستقيم (CRC) أكثر أنواع السرطان شيوعًا بين الرجال، وثالث أكثرها شيوعًا بين النساء. تقدم هذه الرسالة منهجًا قائمًا على البيانات للتنبؤ بمخرجات الإصابة بسرطان القولون والمستقيم من خلال نمذجة ثلاث نتائج سريرية ذات أهمية: حالة السرطان، وتكرار المرض، والحالة الحيوية (البقاء/الوفاة). بالاعتماد على بيانات سريرية من أحد المستشفيات، استكشفت الدراسة النماذج الإحصائية التقليدية، مثل نموذج كوكس لتحليل البقاء والانحدار اللوجستي، إلى جانب تقنيات التعلم الآلي المتقدمة مثل آلات الدعم (SVM)، الغابات العشوائية (RF)، التحسين التدرجي المتطرف (XGBoost)، والشبكات العصبية الاصطناعية (ANN)، لبناء نماذج تنبؤية دقيقة وقابلة للتفسير.

تم استخدام الإحصاءات الوصفية لوصف خصائص المرضى، وتقييم أداء النماذج من خلال التحقق المتقاطع ومؤشرات الأداء مثل المساحة تحت منحنى (AUC) ROC، والحساسية، والنوعية، والقيم التنبؤية. أُجري اختبار المتغيرات في كل من النماذج التقليدية ونماذج التعلم الآلي—باستخدام الاختيار التدرجي في النماذج التقليدية، والإزالة التراجعية ضمن إطار التحقق المتقاطع المتداخل في نماذج التعلم الآلي—وذلك لتعزيز كفاءة النماذج وقابليتها للتفسير. أظهرت نماذج XGBoost و ANN أفضل أداء من حيث التوازن والموثوقية في التنبؤ بحالة السرطان والبقاء، في حين برزت نماذج الغابات العشوائية و XGBoost في التنبؤ بتكرار المرض. كما أكدت نماذج الانحدار اللوجستي ونموذج كوكس أهمية موقع الورم، ومرحلة المرض، وتورط العقد اللمفاوية، وتدخلات العلاج (مثل الجراحة، والعلاج الكيميائي، والعلاج المناعي) عبر جميع المخرجات المدروسة.

وتناولت الدراسة أيضًا، في فصل مستقل، تأثير تحولات السياسات الوطنية لمكافحة السرطان على معدلات انتشار سرطان القولون والمستقيم، ونسبة الإصابات والوفيات الناجمة عنه في المملكة العربية السعودية خلال العقود الثلاثة الماضية. تم استخدام بيانات العبء العالمي للأمراض (GBD 2021) للفترة من ١٩٩٠ إلى ٢٠٢١، وتطبيق نماذج الانحدار السالب الثنائي ومخططات الرقابة الإحصائية القائمة على إحصائية Hotelling T² للكشف عن التغيرات في اتجاهات الإصابة المرتبطة بتغيرات السياسات. وقد تم رصد انحرافات ذات دلالة إحصائية في عامي ١٩٩٥ و ٢٠٢١، تزامنت مع إنشاء السجل السعودي للسرطان، وتوسيع برامج الفحص المبكر لسرطان القولون والمستقيم، على التوالي. ورغم أن الفحص أدى إلى زيادة في معدلات الكشف، إلا أن الاتجاهات المرتبطة بالوفيات أظهرت تحسنًا محدودًا، مما يسلط الضوء على وجود فجوات في فعالية العلاج والحاجة إلى تحسين الرعاية والمتابعة.

وتتمثل الإسهامات الأساسية لهذه الرسالة في تطوير إطار نمذجة متعدد المخرجات، يُقدّم رؤى تتجاوز التنبؤ بنتيجة واحدة فقط. ومن خلال الربط بين التحليل السريري وتقييم الصحة العامة، تهدف هذه الدراسة إلى تعزيز تصنيف المخاطر، ودعم الرعاية الصحية الشخصية، وإثراء جهود مكافحة السرطان الوطنية، مما يساهم في تحسين إدارة سرطان القولون والمستقيم في المملكة العربية السعودية.

1. Introduction

1.1 Background

Chronic illnesses, often referred to as noncommunicable diseases (NCDs), are long-term conditions that develop gradually and are shaped by a mix of genetic, environmental, physiological, and behavioral influences. The main categories of NCDs include cancer, diabetes, cardiovascular issues (e.g., strokes and heart attacks), and respiratory disorders such as asthma and chronic obstructive pulmonary disease (COPD) [1] In KSA, around 35% of all deaths—many occurring before individuals reach 75 years—are linked to these four disease categories. The increasing prevalence of early death, disability, and illness due to NCDs creates substantial obstacles for both healthcare systems and broader economic sustainability [2]

In this context, studies examining chronic disease prevalence among older adults in KSA highlight an alarming pattern: this population bears a significantly higher burden of multiple comorbidities. Major contributing factors include body mass index (BMI), lifestyle habits, and a history of smoking. These results underscore the urgent necessity for holistic health interventions. centered on prevention and disease management—especially for high-risk groups like older adults, who experience the most elevated death rates linked to chronic illnesses [3].

A nationwide study analyzing responses from 24,012 households obtained through the 2018 Household Health Survey explored the rates of chronic illnesses and general mortality in Saudi Arabia. The results revealed a substantial presence of chronic diseases—such as diabetes, high blood pressure, heart conditions, and cancer—and emphasized the critical necessity for public health interventions focused on controlling these conditions, particularly among high-risk and underserved groups [4].

According to a recent data from IARC in 2022, in KSA, CRC ranks as the most commonly diagnosed cancer among men and is the third most prevalent type among women [5]. Considering that CRC is a major contributor to cancer-related illness and death in KSA, prioritizing early detection and preventive measures is of paramount importance.

CRC represents 13.3% of all newly identified cancer diagnoses in KSA, totaling approximately 3,750 cases, and it is responsible for 14.1% of cancer-related deaths nationwide [5]. These statistics show that CRC significantly contributes to the national cancer mortality burden. Despite the presence of a well-developed healthcare system, Approximately 73% of CRC cases in Saudi Arabia are diagnosed at an advanced stage [6], significantly limiting the success of treatment and contributing to higher death rates.

In KSA, the standardized incidence rate of CRC by age has shown a steady upward trend, doubling from 5.0 to 10.5 per 100,000 population between 2001 and 2009 [6]. While CRC predominantly affects individuals over 50 years old, recent studies have identified a troubling increase in incidence among younger populations under the age of 50 years old, suggesting a possible future strain on the healthcare system. These trends highlight the critical importance of implementing effective early detection measures and focused screening initiatives [6].

The upward trend in CRC incidence within KSA is attributed to a combination of lifestyle, environmental, and hereditary factors. Dietary patterns marked by frequent consumption of red meat, sedentary behaviors, and tobacco use have all been linked to this upward trend [6]. Both obesity and diabetes mellitus, which are widely prevalent in the Saudi population, serve as major contributing risk factors. As of 2019, the prevalence of obesity and overweight among adults aged 18 years old and above was recorded at 20.2% and 38.2%, respectively [7]. In addition, rising trends in Type 2 Diabetes Mellitus (T2DM) incidence has increased sharply over time, rising from 23.7% in 2000 to 25.4% in 2009 [8], and reaching a staggering 39.5% in 2022, according to epidemiological modeling [9]. These patterns reinforce the critical necessity for focused public health initiatives. Additionally, inherited susceptibility such as Lynch syndrome has been identified in a portion of CRC cases, emphasizing the value of individualized risk stratification [10].

A comprehensive review of CRC trends in Saudi Arabia between 2001 and 2018, utilizing data from the Saudi Cancer Registry, showed a dramatic 335.6% growth in CRC cases and a 56.4% increase in its share of total cancer diagnoses, rising to 12.2% by 2018. The age-standardized incidence rate (ASR) increased by 152%, with the median age at diagnosis being 60 for males and 58 for females. The highest ASRs were found in the Eastern Region

and Riyadh, highlighting the urgent need to strengthen screening programs, implement effective prevention strategies, and raise public health awareness to combat the growing CRC burden in the Kingdom [11].

Another study explored the demographic and clinical predictors of advanced -stage CRC through a retrospective cohort analysis of 17,541 patients recorded in the Saudi Cancer Registry between 1997 and 2017. Late-stage disease was defined as distant metastases, while early-stage included localized and regional tumors. The analysis found that 29% of patients were diagnosed at an advanced stage, with higher risk observed among individuals under 50 and unmarried women. Variations in risk were also noted across geographic regions and between sexes, and individuals diagnosed after 2001 were more likely to present with advanced-stage disease. These findings underscore the necessity of adopting targeted screening approaches to facilitate earlier diagnosis and enhance clinical outcomes [12] .

An analysis of the epidemiological profile of CRC in Saudi Arabia identified it as the second most prevalent cause of cancer-related mortality in the country. The findings indicated a CRC incidence rate of 14.6% and a mortality rate of 1.48% in 2020. Between 2006 and 2016, the number of CRC cases rose by 8% and 7%, respectively. Factors contributing to this increase included low educational levels, sedentary behavior, and high rates of obesity. The study highlighted the urgent need for comprehensive cancer control strategies focused on primary prevention and raising public awareness to lower CRC incidence and enhance survival outcomes [13].

A study[6] proposed Saudi Arabia as an ideal setting for the adoption of Predictive, Preventive, and Personalized Medicine (PPPM) frameworks, particularly in the context of CRC. The analysis emphasized the increasing CRC burden and identified the absence of a comprehensive national screening program as a major contributor to the high proportion of late-stage diagnoses, with nearly half of the cases presenting with regional or distant metastases. The study recommended integrating early detection efforts, risk-based predictive models, and individualized treatment strategies to enhance clinical outcomes. It also highlighted the importance of lifestyle changes—such as improving dietary habits and increasing physical activity—in alleviating the rising CRC burden. Also, in a separate retrospective evaluation [14] conducted within the Ministry of National Guard–Health

Affairs, late-stage presentation and the absence of surgical or chemotherapy treatment were found to be the strongest predictors of poor survival. These findings underscore the need for a nationwide, population-based screening initiative to support early diagnosis and improve survival prospects.

A continuous rise in CRC incidence has been reported in KSA [15], with the disease now considered the most common cancer among Saudi men. The absence of a national CRC screening initiative has contributed to approximately half of the cases being identified at advanced stages. To address this, the authors advocated for the incorporation of CRC screening within primary healthcare services and the initiation of national campaigns aimed at promoting healthier lifestyle habits. They also anticipated a marked rise in cancer diagnoses by 2040, influenced by demographic transitions, evolving lifestyle patterns, and an aging population.

Detecting CRC at an early stage is vital for enhancing survival outcomes. Although several factors impact patient survival, treatment modality remains a key determinant. Lately, AI and ML have emerged as promising tools for enhancing the accuracy and reliability of chronic disease diagnosis and prediction, including CRC. A growing body of research has explored the application of ML techniques in CRC screening and prevention, underscoring their potential to significantly improve patient prognosis.

A recent investigation conducted in Morocco examined the factors influencing CRC survival using an interpretable machine learning (ML) framework. The model highlighted key prognostic indicators, including demographic characteristics, cancer staging, and treatment modalities, offering meaningful insights into patient outcomes. By facilitating the identification of high-risk cases, the study supports the development of more individualized therapeutic strategies. This approach demonstrates significant potential to strengthen clinical decision-making and improve CRC management and survival rates within the Moroccan healthcare system [16].

Building upon this foundation, the current study underscores the significant promise of advanced ML techniques in identifying high-risk populations and improving CRC outcomes in KSA. This study presents a novel strategy for tackling the ongoing challenges of CRC

management—specifically cancer status, recurrence, and survival—through the development and validation of predictive models that integrate both cutting-edge ML techniques and traditional statistical approaches. By employing this comprehensive analytical strategy, the study aims to advance CRC prevention and patient care by enabling more accurate and individualized risk assessments. Expected outcomes include better identification of high-risk individuals and deeper understanding of disease progression, ultimately supporting targeted interventions and reinforcing national CRC healthcare strategies.

To accomplish these objectives, the study will begin with descriptive statistical analyses to summarize the demographic and clinical profiles of the participants. A suite of ML algorithms—including Support Vector Machines, Random Forests, Extreme Gradient Boosting, and Artificial Neural Networks—will be employed to construct predictive models for three key CRC outcomes - cancer status, recurrence and vital status. For survival analysis, Cox proportional hazards regression will be utilized to capture time-to-event dynamics related to recurrence and overall patient prognosis. Model performance will be assessed using ROC curve analysis, while cross-validation techniques will be implemented to minimize overfitting and enhance the models' generalizability. To optimize predictive accuracy and interpretability, feature selection will be conducted using Sequential Backward Search, systematically removing less informative variables from the models.

Model effectiveness will be evaluated using a comprehensive set of performance metrics, including Area Under the Curve (AUC), sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV), all assessed within a cross-validation framework. The broader significance of this research lies in its potential to enhance risk stratification, enable earlier and more tailored clinical decision-making, and inform long-term strategies to reduce the CRC burden. By integrating clinical expertise with advanced data-driven modeling, this study seeks to deliver a practical and impactful approach to improving cancer outcomes and advancing public health efforts KSA.

1.2 Literature review

CRC is one of the most widespread malignancies globally, posing significant obstacles to prognosis, diagnosis, and treatment. Improving survival outcomes requires early detection,

and ML techniques have become powerful instruments in recent years for increasing the accuracy of CRC diagnosis and prognosis. The World Health Organization (2020) [17], The increasing prevalence of CRC worldwide calls for the creation of novel approaches to enhance screening and diagnostic precision.

Numerous studies have utilized supervised learning methods—such as decision trees, support vector machines, and deep learning—to forecast CRC outcomes based on clinical and genetic information. One study [18] for instance, concentrated specifically on dietary intake variables to estimate CRC risk through ML models. The primary features incorporated in the predictive modeling include:

- I. **Dietary Patterns:** The investigation analyzed various food items and nutrient consumption levels to explore how particular eating behaviors relate to the risk of CRC. Information was obtained using dietary surveys, allowing for an evaluation of food intake patterns and nutrient composition.
- II. **Textual Data Processing:** A novel component of the study involved processing textual descriptions of food ingredients. These were converted into a structured corpus, allowing the identification of key ingredients most associated with CRC risk through frequency and relevance analysis.

To enhance model performance, a two-step feature selection approach was used, incorporating techniques such as logistic regression, Boruta, and Knockoff filters to identify the most significant predictors of CRC risk. This approach reduced the impact of superfluous or unnecessary factors while allowing research to concentrate on the most significant dietary characteristics.

A comprehensive review [19] emphasized the transformative impact of AI in the management of CRC. It highlighted the ability of AI technologies to enhance early detection, diagnosis, treatment planning, and prognosis in CRC care. Through the analysis of large-scale clinical and omics data, ML supports clinicians in identifying high-risk individuals and tailoring personalized treatment strategies. The review reinforces the pivotal role of AI in improving patient outcomes and shaping the future of CRC management.

A systematic review [20] critically examined the potential of ML techniques for predicting CRC risk. It analyzed 14 studies published between 2011 and 2021 that utilized ML models such as random forests, neural networks, and LR. These models were applied to large-scale datasets, with sample sizes ranging from over 17,000 to more than 2.5 million individuals. Although the ML algorithms exhibited strong predictive performance, with AUC values ranging from 0.738 to 0.896, no single model consistently outperformed the others. The review also highlighted several challenges, including limited external validation, variation in algorithms, and complexities in assessing cost-effectiveness. It concluded that while ML shows great promise for CRC risk prediction, robust validation and effective clinical implementation are necessary for broader application.

A detailed comparative analysis [21] was conducted to evaluate the effectiveness of various ML models in predicting mortality and recurrence among CRC patients. The models examined included decision trees, random forests, random survival forests (RSF), gradient boosting, model-based boosting (mboost), and deep learning neural networks (DLNN). The evaluation utilized clinical and demographic data such as age, sex, cancer stage, recurrence status, and treatment details. Notably, gradient boosting and mboost outperformed other approaches, demonstrating higher sensitivity and specificity in forecasting both recurrence and mortality outcomes. In contrast, DLNN demonstrated the weakest predictive performance. These findings highlight the critical role of selecting context-appropriate ML models to enhance predictive accuracy and inform better decision-making in CRC management.

In a study by Liu et al. [22], a range of clinical and pathological factors were identified as significant predictors of CRC recurrence following total mesocolic excision. The analysis was performed using a comprehensive dataset that encompassed variables such as:

I. Preoperative Variables:

- i. Demographics: Gender, age, BMI, and history of alcohol and/or tobacco use.
- ii. Clinical Characteristics: American Society of Anesthesiologists (ASA) score, nutrition risk score (NRS 2002), prior surgeries, illness duration, and previous chemotherapy or radiotherapy.

- iii. Medical History: Conditions such as heart disease, hypertension, diabetes, ileus, and anemia.
- iv. Laboratory Tests: Preoperative levels of albumin, carcinoembryonic antigen (CEA), and carbohydrate antigens CA19-9, CA125, and CA72-4.
- v. Tumor Characteristics: Tumor stage (T-stage, N-stage), size, number, shape, nerve or vascular invasion, and pathological type.

II. Intraoperative Variables:

Type and duration of surgery, amount of intraoperative bleeding, number of lymph nodes excised, surgical approach (e.g., open vs. laparoscopic), and emergency status.

III. Postoperative Variables:

Follow-up laboratory tests (CEA, CA19-9, CA125, CA72-4, procalcitonin, CRP, serum amyloid A [SAA], and neutrophil-to-lymphocyte ratio [NLR]) and tumor recurrence status.

This structured dataset was used to develop ML models, emphasizing the value of both clinical features and biomarkers in predicting recurrence.

Cakmak et al. [23] explored the application of supervised learning models and genetic data to predict susceptibility to CRC. Their predictive modeling focused primarily on single nucleotide polymorphisms (SNPs) associated with immune system regulation and immune-related pathways. Using a patient-level genetic dataset, the study identified associations between specific SNP variants and CRC risk. The findings highlighted the critical role of genetic susceptibility in the development of CRC and showed that combining genomic information with ML algorithms can improve prediction accuracy.

Talebi et al. [24] developed ML models to predict metastasis in rectal cancer patients using clinical data from 1,127 individuals treated at Taleghani Hospital between 2013 and 2019. Metastasis was the primary outcome variable, with independent variables including demographic and clinical features such as tumor stage, tumor size, type of treatment, number of lymph nodes involved, tumor grade, and comorbidities like diabetes mellitus. ML algorithms including neural networks, random forests, and support vector machines were

applied to develop predictive models. The study found that tumor stage, lymph node involvement, and treatment type were the most significant predictors of metastasis.

A review [25] explored the role of ML in improving CRC risk prediction through the use of routinely collected healthcare data. It emphasized the incorporation of a wide range of variables—such as demographic details, clinical backgrounds, and laboratory findings—into predictive modeling. The review underscored the potential of ML to enhance CRC risk stratification by utilizing extensive real-world datasets.

A study [26] examined the use of ML models for detecting CRC through standard laboratory test data. Routine blood markers—including platelet counts, white blood cell counts, and hemoglobin levels—were utilized to construct predictive models. Methods such as logistic regression, random forests, and support vector machines were applied to assess CRC risk. The findings highlighted the potential of incorporating non-invasive laboratory data into predictive models to support earlier diagnosis and minimize dependence on invasive screening methods.

Chen et al. [27] developed a user-friendly CRC prediction model using data from 234,044 Asian adults. The model incorporated a wide range of features, including demographic variables (age, gender), lifestyle behaviors (smoking, alcohol consumption, and diet), and clinical history (family history of CRC and existing medical conditions). ML techniques were applied to analyze these variables with the goal of improving early detection and prevention strategies. The study highlighted the effectiveness of integrating diverse data sources into risk prediction models for CRC.

An article [28] explored the creation of a CRC research database using routinely collected healthcare data, outlining both the methodological approach and its potential contributions. It described the data integration process, highlighting the importance of incorporating demographic, clinical, and treatment-related variables to better understand CRC patterns and outcomes. The discussion also covered challenges such as harmonizing data from different sources and ensuring robust data governance to protect patient confidentiality. The study illustrated how well-structured databases can facilitate advanced research, strengthen early detection strategies, and improve the quality of CRC care.

The authors also discuss the difficulties involved in harmonizing data from multiple sources and emphasize the importance of strong data governance frameworks to protect patient confidentiality. The study demonstrates that a well-integrated database can support advanced research efforts, enhance early detection initiatives, and ultimately contribute to improved patient management and CRC outcomes.

A study [29] utilized AI techniques to create a predictive model for tumor relapse in patients with stage II–III CRC after curative surgery. Among the models assessed, logistic regression achieved the best predictive outcome with an AUC of 0.678. In addition, genomic profiling revealed frequent occurrences of PIK3CA and DNMT3A mutations in older patients, indicating age-related variations in tumor behavior. These results highlight the importance of incorporating molecular and genetic characteristics into AI-driven prediction models to improve CRC patient care.

One study [30] investigated the use of ML methods to forecast recurrence following surgery in patients with stage IV CRC patients. The study analyzed data from 999 individuals, with an 80/20 split for training and testing, respectively. Four ML techniques—logistic regression, decision trees, Gradient Boosting, and LightGBM—were assessed. Gradient Boosting achieved superior performance, attaining an AUC score of 0.881 and an F1 score of 0.912 in the training set. Although the logistic regression model showed the lowest AUC (0.692) in the test set, Gradient Boosting retained solid predictive power with an AUC of 0.734. The most influential predictors of recurrence included chemotherapy, patient age, LogCEA, CEA levels, and anesthesia duration. These findings affirm the utility of ML models, particularly Gradient Boosting approaches, in forecasting tumor recurrence and improving clinical decision-making for stage IV CRC patients.

In summary, existing literature highlights the transformative potential of AI and ML in advancing the management, detection, and prediction of CRC. Numerous studies have demonstrated that ML algorithms can improve early detection, accurately identify high-risk individuals, and predict recurrence, thereby enabling more personalized interventions and better clinical outcomes. As research in this field continues to grow, the emergence of accessible, AI-driven tools holds significant promise for further revolutionizing CRC care and contributing to improved health outcomes across affected populations.

1.3 Problem Description

This thesis aims to address the current limitations in integrated, evidence-based approach for accurately predicting and managing key CRC outcomes; and data-driven approach to predicting and managing multiple CRC outcomes - specifically cancer status, recurrence, and vital status - within KSA. Despite ongoing improvements in healthcare infrastructure, CRC continues to be a major contributor to cancer-related illness and death in KSA, with late-stage diagnoses and limited outcome-specific predictive tools contributing to suboptimal clinical decision-making and long-term health outcomes.

To tackle this challenge, we adopt a multi-method strategy that combines advanced ML techniques with conventional statistical models (Cox proportional hazards and logistic regression) to enhance prediction accuracy, interpretability, and clinical relevance. In addition, a separate component of the study investigates the impact of cancer control policy transitions on national trends in CRC prevalence, incidence, and mortality.

To guide our investigation, the following focused research questions have been formulated:

- I. Risk Factors Identification: What are the potential demographic, clinical, and behavioral risk factors within the KSA population that can be leveraged to effectively predict CRC recurrence, cancer status, and vital status?
- II. Modeling Techniques: Which traditional statistical methods and ML algorithms are most effective for modeling these CRC outcomes in terms of accuracy, robustness, and clinical interpretability?
- III. Model Validation and Performance Analysis: How can the developed models be rigorously evaluated in terms of predictive accuracy, reliability, and generalizability? Which performance metrics and validation strategies are most effective in ensuring clinically meaningful outcomes?
- IV. Healthcare Implications: What are the practical implications of the identified clinical risk factors and predictive model outputs for improving personalized CRC care, guiding outcome-based interventions?
- V. Policy Impact Assessment: How have historical transitions in national cancer control policies influenced CRC trends (prevalence, incidence, mortality) over time in KSA?

Together, these questions define the central objectives of this study, which aims to advance CRC outcome prediction, inform personalized care, and support evidence-based public health strategies in KSA.

1.4 Objective

This thesis aims to address the critical issue of CRC management by developing predictive models for three key clinical outcomes: cancer status, recurrence, and vital status. The study will utilize a comprehensive clinical dataset obtained from King Fahd Specialist Hospital in Dammam, combined with population-level data from the GBD 2021 report. By applying both advanced ML algorithms and traditional statistical methods, the research seeks to improve predictive accuracy, support personalized care, and evaluate the long-term impact of cancer control policies in KSA.

The following are the primary objectives of this thesis:

- I. **Risk Factors Identification:** To identify demographic, lifestyle, and clinical risk factors associated with CRC cancer status, recurrence, and vital status among the Saudi population. This objective aims to evaluate how these factors can be effectively integrated into predictive models to improve accuracy, enhance interpretability, and support clinical relevance in outcome prediction.
- II. **Survival Analysis and ML Model Development:** To develop robust predictive models for the three CRC outcomes using both statistical methods and ML algorithms. This study will assess numerous ML techniques, such as support vector machines (SVM), random forests (RF) to identify the most effective methods for predicting survival outcomes.
- III. **Policy Impact Analysis:** To assess the impact of transitions in national cancer control policies on CRC prevalence, incidence, and mortality in KSA between 1990 and 2021 using GBD data. This will be achieved through negative binomial regression and Hotelling's T^2 -based control charts, enabling the detection of statistically significant shifts over time.

Through these objectives, the study aims to contribute novel insights into CRC prediction and monitoring, improve outcome-based decision-making, and support evidence-driven health policy in KSA.

1.5 Methodology

This thesis employs a comprehensive and innovative methodology to address the multifaceted challenges associated with CRC outcome prediction in KSA. Moving beyond recurrence alone, this research models three key clinical outcomes: cancer status, recurrence, and vital status. To achieve this, both advanced ML techniques and conventional statistical techniques - notably Cox proportional hazards regression and logistic regression - are applied. The methodology is organized into five primary components:

I. Risk Factor Identification

A broad set of clinical, demographic, and treatment-related variables from hospital-based patient records will be considered as candidate predictors for CRC outcomes. These variables are grouped into the following domains:

- i. **Demographic Variables:** Patient sex, age at diagnosis, and marital status.
- ii. **Tumor Pathology and Staging:** Primary tumor site and histology (both ICD-O-3 coded), summary stage classification, and regional lymph node involvement categorized as: (i) regional nodes positive, (ii) regional nodes negative, or (iii) no nodes examined.
- iii. **Treatment Modalities Received:** Information on surgical intervention, chemotherapy, radiotherapy, immunotherapy, and palliative care.
- iv. **Follow-Up and Time-to-Event Variables:** Date of CRC diagnosis, date of recurrence detection, date of last contact or death, and date of last cancer status evaluation.
- v. **Outcome Variables:** cancer status, recurrence status, and Vital status are The primary outcomes of interest in this thesis.

II. Statistical Analysis and ML for Outcome Prediction

This phase centers on building predictive models for the three CRC outcomes using a combination of traditional statistical methods and ML algorithms.

- i. **Descriptive Statistics:** A summary analysis will be performed to describe the clinical and demographic characteristics of the participants, offering context for the analyses that follow.
- ii. **Traditional Statistical Models:** Predictive models will be developed using Cox proportional hazards regression for time-to-event outcomes, and logistic regression (LR) for binary classifications of cancer status, recurrence and vital status.
- iii. **ML Algorithms:** Predictive models will be developed using RF, Support Vector Machines with both linear and radial kernels (svmLinear and svmRadial), XGBoost, and ANN.

III. Validation and Robustness of Models

To ensure the strength and dependability of our results, a comprehensive validation strategy will be applied that extends beyond conventional model evaluation techniques:

- i. **Cross-Validation:** Each model will be subjected to cross-validation on the training data to minimize overfitting and promote generalizability across various patient groups.
- ii. **Sensitivity Analysis:** Model stability will be assessed through sensitivity analyses, where key predictor variables are systematically varied and resulting changes in performance metrics are examined. This method will help evaluate the influence of individual risk factors and ensure the model remains robust across diverse clinical scenarios.

IV. Feature Selection:

To optimize model performance, enhance interpretability, and reduce overfitting, feature selection will be applied at both traditional and ML modeling stages. In the traditional statistical models (Logistic and Cox regression), a stepwise selection strategy will be used to identify significant predictors based on model fit criteria (e.g., AIC). For ML models, Sequential Backward Elimination will be employed within a nested cross-validation framework, starting with all candidate features and iteratively removing the least informative variables. This embedded feature selection process

ensures that the final models are both parsimonious and predictive, tailored to each clinical outcome—cancer status, recurrence, and survival..

V. Policy Impact Analysis on CRC Trends

In addition to individual-level modeling, this thesis includes a population-level assessment of the impact of national cancer control policies on CRC trends in KSA. This analysis provides a broader public health context to complement the clinical prediction components of the study.

Using data from the GBD 2021 study, covering the period 1990 to 2021, the analysis will:

- i. Employ negative binomial regression to estimate age- and sex-adjusted annual rates of CRC prevalence, incidence, and mortality.
- ii. Apply Hotelling's T^2 -based control charts to detect statistically significant deviations in these rates, corresponding to possible shifts in national cancer policy or intervention programs.

This component aims to evaluate the effectiveness of KSA's cancer control strategies and connect policy transitions to changes in disease burden over time. It reinforces the multi-level nature of this research by linking predictive modeling with health system performance evaluation.

VI. Software and Computational Feasibility

All data processing, modeling, and analytical procedures in this study were carried out using the R programming environment. The clinical and population-level datasets used for this research were provided in Microsoft Excel format and were imported and exported using packages such as `readxl`, `openxlsx`, and `writexl`, which supported efficient handling of structured spreadsheet data.

A diverse set of R libraries was utilized to implement ML algorithms, perform survival and regression modeling, and carry out statistical validation and control chart analysis. Key packages for ML included `caret` for unified modeling workflows and resampling, `xgboost` for gradient boosting, `randomForestSRC` for RF survival

models, C50 and bst for decision tree and boosting models, and themis for addressing class imbalance through techniques like Synthetic Minority Over-sampling Technique (SMOTE).

For traditional statistical modeling and diagnostics, the study employed survival, MASS, car, matlib, and lmtest, supporting generalized linear models, Cox regression, and multicollinearity checks. Survival curve visualization and assumption checking were facilitated through survminer, while competing risks modeling was performed using cmprsk.

Model performance was evaluated using ROCR, pROC, PRROC, and caTools, enabling robust computation of ROC curves, AUC scores, and other classification metrics.

To evaluate the impact of national cancer control policies, statistical process control and multivariate quality control techniques were applied using Hotelling, MSQC, qcc, and spc. These libraries supported the construction of Hotelling's T^2 control charts and phase-wise monitoring of trends in CRC prevalence, incidence, and mortality.

Collectively, this software ecosystem ensured reproducibility, computational efficiency, and analytical rigor throughout all stages of data management, modeling, and interpretation.

VII. Expected Outcomes:

This novel methodological approach is anticipated to improve the detection of high-risk individuals and offer enhanced insights into critical clinical outcomes in CRC, including recurrence, cancer status, and vital status. Through the integration of advanced ML algorithms and conventional statistical techniques, the study seeks to produce precise, interpretable, and clinically actionable predictions that facilitate personalized patient care. Furthermore, the policy-oriented aspect of the research is expected to generate meaningful evidence on the effectiveness of national cancer control initiatives in KSA by examining how historical policy changes have shaped CRC trends over time.

Collectively, these outcomes will contribute to more informed, data-driven healthcare strategies for CRC prevention, early identification and intervention. Ultimately, the study aspires to transform patient care and public health planning in KSA through robust predictive tools and actionable insight

1.6 Unique Contributions and Innovations

This research introduces several unique contributions to the field of CRC outcome prediction and survival analysis, distinguishing it from existing studies and enhancing its relevance to CRC management and public health planning in KSA:

- i. **Multi-Outcome Prediction Framework:** While many studies emphasize CRC recurrence as a primary outcome, this thesis extends the scope by modeling three clinically significant outcomes - cancer status, recurrence, and vital status to offer a more thorough and useful comprehension of survivorship, treatment response, and illness progression.
- ii. **Statistical and ML Techniques:** This study employs both traditional statistical methods and advanced ML algorithms to enable rigorous comparison and validation across modeling paradigms, and to offer both predictive power and clinical interpretability.
- iii. **Innovation in ML Application:** By utilizing techniques like Sequential Backward Search for feature selection, the study enhances the transparency and clinical utility of ML models applied to CRC outcomes.
- iv. **Performance Over Standard Models:** The predictive models developed aim to outperform conventional approaches in terms of accuracy and robustness, as evaluated by metrics such as AUC, sensitivity, specificity, PPV, and NPV. Optimized ML algorithms - such as SVM, RF, and ensemble methods—are expected to deliver improved outcome predictions, especially for high-risk CRC patients.
- v. **Policy Impact Analysis:** A novel contribution of this thesis is its dedicated policy analysis chapter, which evaluates how transitions in national cancer control policies have influenced CRC prevalence, incidence, and mortality over time in

KSA. This public health perspective links individual-level modeling to broader systemic efforts in cancer control.

Together, these innovations position the study to contribute meaningfully to both the clinical and policy dimensions of CRC prevention and management in KSA.

2. Machine Learning Framework and Methodological Foundations

This section provides a structured overview of the key ML concepts and methodologies employed in this thesis. It focuses on various algorithms, statistical models, and evaluation metrics essential for developing accurate and interpretable predictive models for CRC outcomes. It also addresses the balance between model complexity, generalization, and predictive performance, ensuring the optimal selection of techniques suited to the study's objectives.

2.1 Machine Learning Techniques

The application of ML techniques in healthcare research continues to expand, offering robust methodologies for analyzing complex datasets and enhancing predictive accuracy[31]. These advanced computational methods facilitate the discovery of intricate patterns within large-scale health data, thereby improving predictions, model accuracy, and addressing data limitations such as missing values[32]. This chapter provides an outline of the specific ML algorithms employed in this thesis, emphasizing their application to CRC prediction, risk factor analysis, and survival modeling within the context of Saudi Arabia.

ML, a leading branch of AI, entails creating algorithms that can derive insights and generate predictions or decisions from data. By leveraging statistical, probabilistic, and optimization techniques, ML systems enhance their accuracy through repeated learning cycles without being explicitly programmed for every task [33]: In the realm of biomedical research and healthcare, ML holds substantial promise for extracting valuable insights from multifaceted datasets, aiding in early detection, diagnosis, risk factor analysis, and prognostic modeling [34], [35]

ML techniques are generally categorized into four broad paradigms: supervised, unsupervised, semi-supervised, and reinforcement learning.

I. Supervised Learning

This approach uses labeled datasets, where input variables (features) and the labels that correlate to them are known, to train models. The model learns to establish relationships between inputs and outputs, allowing for precise forecasts on invisible data. A variety of ML techniques—including LR, Decision Trees (DT), Random Forests (RF), Support Vector Machines (SVM), and Neural Networks (ANN)—are frequently employed in healthcare settings to perform risk stratification and outcome estimation. [34]

II. Unsupervised Learning

In contrast to supervised learning, unsupervised learning operates on datasets that lack predefined labels. Instead of relying on known outcomes, these algorithms aim to uncover hidden patterns or intrinsic structures within the data. Techniques such as K-means and Hierarchical Clustering are commonly utilized for grouping similar observations, while methods like Principal Component Analysis (PCA) are employed for reducing dimensionality. These approaches are particularly useful in descriptive analytics and the identification of unusual or outlier behavior within datasets[33]

III. Semi-Supervised Learning

This approach serves as a middle ground between supervised and unsupervised learning by leveraging both labeled and unlabeled data. Semi-supervised learning is especially beneficial in situations where obtaining labeled data is expensive or labor-intensive, allowing models to achieve better performance using a relatively small set of labeled examples [36]

IV. Reinforcement Learning

Unlike other learning paradigms, reinforcement learning focuses on teaching an agent to make a series of decisions by engaging directly with its environment. Through a system of rewards and penalties, the agent progressively learns to choose actions that maximize long-term benefits. This method is particularly effective in dynamic, uncertain settings where labeled training data is limited or unavailable [37]

In this study, supervised learning techniques are utilized to model CRC outcomes by training algorithms on labeled clinical datasets from King Fahd Specialist Hospital in Dammam. The primary focus is to predict cancer status, recurrence, and vital status using demographic, clinical, and treatment-related variables. Performance metrics such as AUC, sensitivity, specificity, PPV, and NPV are used to evaluate model accuracy and reliability. Feature selection is performed using Sequential Backward Search to enhance model efficiency and interpretability.

Outlined below is a summary of the ML techniques employed in this thesis:

2.1.1 Logistic Regression (LR)

LR is a supervised learning algorithm designed to address problems involving two-class classification, where the target variable takes on one of two distinct categories (e.g., 0 or 1, positive or negative). Although called “regression,” it is fundamentally a classification algorithm because it models the probability that a given observation is associated with a specific class label [38].

The core of LR is the sigmoid function, which transforms continuous input values into a range between 0 and 1, effectively representing probabilities [39]:

$$g(z) = \frac{1}{1 + e^{-z}}$$

Here:

- z is a linear combination of inputs, $z = \theta^T x$
- The output is a probability score that can be interpreted as the likelihood of a particular class.

LR estimates the likelihood of $y=1$ given inputs x through the sigmoid activation function [39]:

$$P(y = 1 | x; \theta) = \frac{1}{1 + e^{-\theta^T x}} = g(\theta^T x)$$

Here, θ is a vector of model parameters, and the output is the probability of belonging to class 1. This setup assumes that the dependent variable follows a Bernoulli distribution.

LR expresses the natural logarithm of the chances of the outcome as a linear function of the input variables [33]:

$$\text{logit}(p) = \log \left(\frac{p}{1-p} \right) = \theta^T x$$

Taking the exponential of both sides gives the odds[33]:

$$\frac{p}{1-p} = e^{\theta^T x}$$

This formulation demonstrates that LR models a linear relationship between the predictors and the log chances of the outcome.

The model coefficients θ are determined through Maximum Likelihood Estimation (MLE). For a dataset of observations (x_i, y_i) , the likelihood function is[40]:

$$L(\theta) = \prod_{i=1}^n p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}$$

Taking the natural logarithm gives the log-likelihood function[40]:

$$l(\theta) = \sum_{i=1}^n [y_i \log(p(x_i)) + (1 - y_i) \log(1 - p(x_i))]$$

This function is maximized using numerical optimization techniques such as gradient ascent or Newton-Raphson.

Advantages of Logistic Regression

- I. Probabilistic Interpretation: Provides probability estimates for each class, which can be useful in decision-making [41].
- II. Ease of Implementation: Simple to implement and computationally efficient, For instance, a study comparing various ML models for breast cancer diagnosis found that LR not only outperformed more complex models like DT and SVM but also demonstrated significant computational efficiency, rendering it especially appropriate for use in medical contexts [42].
- III. Interpretability: Coefficients are straightforward to interpret, especially when features are not correlated[43].

Limitations of Logistic Regression

- I. Assumes that the predictors have a linear association with the log-odds of the dependent variable [44].
- II. Not suitable for complex, non-linear decision boundaries unless transformed[45].
- III. Sensitive to multicollinearity among predictors[46].

2.1.2 Decision Trees (DT)

DT is a supervised learning technique applied to both classification and regression issues. DT are widely used ML models that provide excellent interpretability and robust predictive performance. They are applicable to both categorical and continuous outcome prediction tasks and do not require domain-specific knowledge or parameter tuning, making them straightforward and easy to understand[47]

Structure of a DT

DT is structured with leaves at the base, resulting in an inverted tree-like appearance. Its components include[48], [49]:

- I. Target or Class Variable: Represents the outcome being predicted, such as disease status or risk level.
- II. Root Nodes: The starting point of the splitting process, determined by the feature that best separates the classes.
- III. Branches: Connections between nodes, representing decisions made based on the feature's values.
- IV. Pure/Impure Nodes: A node is pure if it contains only one class, and impure if it has a mix of classes.
- V. Decision Nodes: Impure nodes that necessitate further splitting based on a different feature.

VI. Leaf Nodes (Terminal Nodes): Pure nodes that signify the end of a branch where predictions are made.

This top-down, greedy approach iteratively selects the most suitable attribute at every node from the root to the final terminal node, ensuring the dataset is effectively partitioned at each step [33].

Attribute or Feature Selection

Effective feature selection is critical for constructing a robust decision tree. Several metrics are commonly used to determine the best attribute for splitting:

- **Entropy:** Quantifies the level of disorder or uncertainty within a dataset. A node with mixed composition is considered to have higher entropy than a pure node. It is calculated as:

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

Where P_i represents the probability of randomly choosing an instance that belongs to class i , and S is a subset of the training examples [33]

- **Information Gain:** Quantifies the decrease in entropy caused by splitting the dataset based on a specific property. It's calculated with the following formula [33]:

$$G(S, A) = H(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} H(S_v)$$

Where S represents the dataset, and S_v is the subset of data where attribute takes the value v

- **Gini Index:** Measures the impurity of a node by calculating the probability of misclassification. It is computed as[33]:

$$\text{Gini}(D) = 1 - \sum_{i=1}^C P_i^2$$

Where C denotes the total number of classes, and P_i is the proportion of samples belonging to class i .

- **Gain Ratio (C4.5):** An improvement over information gain that adjusts for bias by normalizing the information gain using the intrinsic information of a split[50].
- **Chi-Square Feature Selection and Integration with C5.0 Algorithm :** The Chi-Square test evaluates attribute significance by measuring dependence on the target

variable. While the C5.0 algorithm, developed by Ross Quinlan, improves upon C4.5 with boosting and pruning, it doesn't use Chi-Square tests inherently. However, integrating Chi-Square feature selection with C5.0 improves accuracy, robustness, and reduces overfitting [51], [52].

DT Algorithms

There are several widely used algorithms for constructing decision trees:

- I. CART (Classification and Regression Trees): Developed by Breiman et al, CART builds binary decision trees by employing the Gini index for classification tasks and Mean Squared Error for regression [53]
- II. C4.5: Developed by Quinlan, this algorithm improves upon ID3 by handling continuous and categorical data, using gain ratio for feature selection, and incorporating pruning for better generalization [54].
- III. C5.0: An enhancement of C4.5 with faster performance, better memory efficiency, and support for boosting. It is commonly used for large-scale classification tasks [55].
- IV. CHAID (Chi-Squared Automatic Interaction Detection): this algorithm is particularly effective for categorical data, using statistical tests to determine splits [56]
- V. QUEST (Quick, Unbiased, Efficient, Statistical Tree): Developed by Loh, QUEST is designed for computational efficiency and unbiased splitting of continuous and categorical data [57]

Advantages of DT

DT provide various benefits that make them a popular choice for ML applications. One of their primary strengths is interpretability; decision trees are easy to understand and visualize through graphical representation. This transparency is valuable when communicating results to stakeholders or interpreting the underlying relationships between features. Additionally, decision trees excel at handling nonlinear relationships. They can effectively capture complex decision boundaries without the need for feature scaling or transformation, which is a common requirement for other algorithms such as LR or SVM. Moreover, DT provide significant versatility, as they are applicable to both categorical and continuous outcome prediction tasks. This broad applicability is further enhanced by their simplicity, as they do not require extensive data preprocessing or parameter tuning [33]

Limitations of DT

Although DT have numerous strengths, they also possess significant drawbacks that can affect their performance and generalizability. A significant drawback is their tendency to overfit, particularly when deep trees are constructed with numerous branches. Such models can memorize the training data rather than generalize well to new, unknown input. Additionally, decision trees exhibit instability, Meaning that tiny modifications in the

training data can result in significantly diverse tree architectures. This sensitivity can complicate model interpretation and reliability. Another limitation is the potential for bias, especially when the dataset is imbalanced or when features have differing scales or magnitudes. Furthermore, decision trees may exhibit limited generalization capability, particularly when applied to noisy or complex datasets, where their hierarchical splitting approach may struggle to capture subtle patterns or interactions [33]

Applications of Decision Trees

Decision Trees have been widely adopted in healthcare for disease diagnosis, treatment planning, and risk assessment, offering transparent and reliable support for clinical decision-making. Beyond healthcare, they are also applied in finance and marketing for tasks such as credit scoring, fraud detection, and customer behavior analysis[33].

2.1.3 Random Forest (RF)

RF is a prominent ensemble learning algorithm in ML, renowned for its versatility and robustness in regression and classification tasks. During training, it creates several decision trees and returns the mode of the classes (for classification) or the mean prediction (for regression) of the individual trees [58].

The RF algorithm works by generating many decision trees, each trained on a random portion of the training data using a technique known as bootstrapping. Furthermore, at each split in the tree building, a random subset of features is chosen, increasing tree variety and lowering the danger of overfitting. This randomization ensures that the ensemble model catches a variety of patterns in the data [33]

Advantages of RF

One of the most notable benefits of RF is its high accuracy. By combining predictions from numerous decision trees, RF achieves improved accuracy compared to individual trees. This enhanced accuracy results from the diversity and averaging of predictions, which effectively reduces the overall variance of the model and improves its generalization ability [58]. Additionally, RF demonstrates robustness to overfitting due to the combination of bootstrapping and feature randomness. The algorithm constructs decision trees by drawing random samples of the data with replacement and choosing random subsets of features at each node split. This method helps minimize correlations between the individual trees, further enhancing robustness and performance, particularly in complex datasets [59].

Another advantage of RF is its ability to provide estimates of feature importance. Throughout the training procedure, the algorithm calculates the significance of each feature by determining the decrease in the Gini index or residual sum of squares (RSS) when splits are made using that feature. By averaging these decreases over all trees, RF generates an importance ranking, which aids in feature selection and provides insights into the data's basic

structure [33], [60]. Furthermore, RF is effective at handling missing data. The algorithm can retain accuracy even when a large piece of the data is absent by utilizing proximities between observations to estimate missing values, thus making it robust and reliable in real-world scenarios[59].

Limitations of RF

Despite its obvious advantages, RF does have certain limits. One of the primary disadvantages is It is computationally complex. Training many decision trees can be computationally demanding, especially when working with big datasets. This issue may result in longer training times and increased resource consumption, especially when the algorithm is applied to high-dimensional data[59]. Furthermore, while individual decision trees are simple to read, RF's ensemble structure makes it difficult to understand how each tree contributes to the final forecast. This reduced interpretability can be a drawback when model transparency is required[59]. Furthermore, RF may exhibit bias when handling categorical variables with different numbers of levels. Variables with more categories tend to have more splits, influencing the prediction more heavily and potentially leading to biased results [59].

Applications of RF

RF is frequently used in numerous fields such as healthcare, environmental monitoring, and remote sensing. Its robustness and ability to handle complex data make it suitable for diverse applications [59]

Recent improvements in RF include methods to enhance accuracy and robustness by optimizing decision tree construction, adjusting the feature selection process, and integrating optimization techniques such as Elastic Net and SVM Recursive Feature Elimination. Additionally, attempts have been made to improve the management of high-dimensional datasets and boost their interpretability [61]

2.1.4 Support Vector Machines (SVM)

SVM have been commonly acknowledged as highly effective classifiers, demonstrating their effectiveness across various domains, including applications in the medical field[62]

SVMs are powerful tools for supervised learning tasks encompassing regression and classification. By design, SVMs are tailored for binary classification and are distinguished by their exceptional predictive accuracy [33]

Finding the best separating hyperplane that optimizes the margin between two classes is the basic idea of SVM. Support vectors, or the data points nearest to the decision border, define this hyperplane[63]

Given a training dataset $D = \{(x_i, y_i)\}_{i=1}^n$, where x_i is a feature vector of p dimension and y_i is the label corresponding to x_i with $y_i \in \{+1, -1\}$. The objective is to identify a hyperplane that possesses the following properties:

$$w^T x + b = 0$$

where w denotes the weight vector and b represents the bias term.

Maximizing the difference between the two classes is the goal, which is equivalent to minimizing:

$$\min_{w,b} \frac{1}{2} \|w\|^2$$

Subject to the constraint:

$$y_i(w^T x_i + b) \geq 1, \quad \forall i$$

The Hard Margin SVM is the name given to the aforementioned formulation. Since perfect separation is frequently impossible in real-world situations, slack variables ξ_i are included to a Soft Margin SVM:

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i$$

Subject to:

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

where the trade-off between maximizing the margin and minimizing classification errors is controlled by the regularization parameter C .

Separating Hyperplane

The decision boundary that separates the classes in the feature space is represented by the separating hyperplane, a key idea in SVMs. It is defined by the following equation:

$$w^T x + b = 0$$

Where:

- w is the normal vector that is orthogonal to the hyperplane.

- b is the bias term that defines the hyperplane's displacement from the origin.

The margin, which SVMs aim to maximize, denotes the gap between the hyperplane and the closest instances from each class. We call these closest data points "support vectors". A point's distance to the hyperplane is calculated as:

$$\text{Distance} = \frac{|w^T x + b|}{\|w\|}$$

Maximizing the margin is equivalent to minimizing the norm of the weight vector $\|w\|$. which promotes a robust decision boundary that generalizes effectively to new, unseen data.

Kernel Trick

SVMs can be adapted for non-linear classification tasks by mapping the original input space into a higher-dimensional feature space through the use of a kernel function [64], [65]

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$$

Common kernels include[65], [66]:

- Linear Kernel: $K(x_i, x_j) = x_i^T x_j$
- Polynomial Kernel: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d$
- Radial Basis Function (RBF): $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$

Advantages of SVM

SVMs offer several advantages that make them highly suitable for healthcare applications. Their strong generalization ability and robustness against overfitting are particularly beneficial when working with high-dimensional clinical datasets, where the number of features often exceeds the number of samples. The principle of Structural Risk Minimization enables SVMs to achieve high accuracy in both linear and non-linear classification tasks, which is critical for medical diagnosis, risk prediction, and disease classification [67], [68]. Moreover, their versatility through kernel functions allows SVMs to handle complex biomedical data structures by transforming them into higher-dimensional spaces where a linear decision boundary becomes feasible. Their efficiency in high-dimensional spaces, memory-efficient predictions using support vectors, and proven success in biomedical domains such as genomics and medical image analysis further enhance their value in healthcare settings [37].

Beyond healthcare, SVMs are useful in fields like image and text classification, bioinformatics, and in hybrid approaches integrated with deep learning for improved performance in complex real-world tasks.

Limitations of SVM

SVMs have several notable limitations. They require high computational resources for large datasets, as the training process involves solving a complex quadratic optimization problem. Their performance is highly sensitive to parameter tuning, particularly the choice of kernel and associated parameters, which demands extensive cross-validation. SVMs also tend to exhibit weak performance on imbalanced datasets [69], often favoring the majority class, and they are sensitive to noise and overlapping classes, which can distort the decision boundary. Additionally, SVMs lack probabilistic output, offering only binary predictions without confidence levels, which can be a drawback in applications like healthcare. Moreover, since SVMs are inherently binary classifiers, extending them to multi-class problems increases computational complexity. Finally, selecting and optimizing the appropriate kernel function can be challenging, especially in large-scale problems [67], [68].

Applications of SVM

SVMs have shown strong potential in medical and healthcare applications. They are frequently used in disease diagnosis and outcome prediction, including cancer classification, cardiovascular disease detection, and leukemia incidence and mortality forecasting [39]. Their effectiveness extends to bioinformatics tasks such as protein classification, DNA classification, and protein-protein interaction prediction, where their ability to handle high-dimensional and complex biological data offers clear advantages over traditional statistical methods [37].

Outside the healthcare domain, SVMs are widely used in text categorization, image classification, financial analysis, speech recognition, and predictive control, offering robust performance across diverse, high-dimensional datasets [67].

2.1.5 Artificial Neural Networks (ANNs)

ANNs are computational models Motivated by the biological neural networks seen in the human brain. They are made up of layers of artificial neurons (nodes) that are connected. That handle information by modifying connection weights using learning algorithms. The resurgence of interest in ANNs occurred with the development of backpropagation algorithms and the advent of deep learning architectures, which facilitated more efficient training of multi-layer networks [34]

Structure of ANNs

ANNs generally consist of three types of layers:

- I. Input Layer: Takes in unprocessed data and sends it to the network.
- II. Hidden Layers: Intermediate layers where computations are performed. Multiple hidden layers form deep networks, which allow learning of complex patterns [34]
- III. Output Layer: Generates the end product, such as classification or forecast.

In order to introduce non-linearity, each neuron processes inputs by first applying a mathematical function to the weighted sum of inputs, and then an activation function (such as Sigmoid or ReLU) [70]

Characteristics of ANNs

ANNs possess certain characteristics that distinguish them from traditional statistical models and allow them to pick up intricate patterns in data. These characteristics include adaptability, fault tolerance, non-linearity, and parallel processing [34], [71]

Components of ANNs

ANNs are composed of several critical components, including[34]:

- I. Neuron: These are the fundamental units of computation that get inputs, process them, and then generate outputs. They are commonly represented by a weighted sum of inputs with non-linearity introduced by an activation function.
- II. Activation Functions: These are mathematical operations that add non-linearity to neuronal output. Examples include ReLU, Sigmoid, tanh, etc. They are necessary for neural networks to be able to recognize intricate patterns.
- III. Feedforward Networks (Multilayer Perceptrons - MLPs): Described as networks where information flows through layers of neurons without feedback connections. They are the simplest form of deep learning models.
- IV. Feedforward Networks vs Feedback Networks: Feedforward networks, such as Multilayer Perceptrons (MLPs) and Convolutional Neural Networks (CNNs), allow data to move without cycles or loops in a single direction, from input to output [42]. In contrast, feedback networks, include Long Short-Term Memory Networks (LSTMs) and Recurrent Neural Networks (RNNs), which are ideal for sequential and time-dependent data because they have loops that allow the network to remember prior inputs [72]
- V. Differentiability Requirement: Activation functions in ANN must be differentiable to enable the application of gradient-based optimization algorithms such as backpropagation. This requirement ensures the network can effectively learn through iterative optimization processes. According to Goodfellow, Bengio, and Courville (2017), activation functions like ReLU are particularly effective because their near-linear behavior makes them easier to optimize using gradient-based methods, especially when the network is active and the derivatives remain large[34]

VI. Signal Processing and Thresholding: Neurons fire (activate) when a specific threshold is reached by the weighted sum of the inputs [73]

Types of Artificial Neural Networks

ANNs can be divided into a number of categories according to their design and method of learning. Among them are:

- I. Perceptron: The simplest ANN model consisting of a single layer of neurons used for binary classification tasks [34]
- II. Multi-Layer Perceptron (MLP): a backpropagation-trained network with one or more hidden layers. More complicated tasks like pattern recognition and regression are handled by MLPs [70].
- III. Convolutional Neural Networks (CNNs): primarily made for jobs involving computer vision and image processing. CNNs have shown remarkable efficacy in object identification, segmentation, and image classification [70].
- IV. Recurrent Neural Networks (RNNs): efficient when processing data in a sequential manner. For time-series forecasting and natural language processing, two common variations are Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) [74]
- V. Generative Adversarial Networks (GANs): GANs were first presented by Goodfellow et al. and are used for unsupervised learning and data production tasks. They are composed of two competing networks, a discriminator and a generator [34]
- VI. Transformers: Emerging as a powerful architecture for NLP tasks, Transformers have outperformed RNNs in various language modeling and translation tasks[75]

Advantages

ANNs offer several advantages that make them highly effective in a variety of applications. One of their key strengths is adaptability—their capacity to use training to uncover intricate links and patterns in data, which enables them to tackle diverse and nonlinear problems [34]

ANNs also support parallel processing, allowing them to handle large volumes of data simultaneously and improve computational efficiency, especially in high-dimensional tasks[76]. Additionally, their distributed architecture provides fault tolerance, making them sturdy enough to noise and errors in the input data, that is particularly beneficial in real-world, imperfect datasets[71].

Disadvantages

Despite their strengths, ANNs have some disadvantages. They are particularly prone to overfitting, which might restrict their capacity to generalize to new data, especially when trained on noisy or limited datasets [77]. Additionally, the training of deep neural networks is computationally intensive, often requiring significant processing power and high-

performance hardware, making them less accessible for resource-constrained environments [78]. Another major limitation is their lack of interpretability; ANNs are often considered "black-box" models because it is challenging to comprehend and describe their own decision-making processes, which can be problematic in fields where transparency is critical, such as healthcare [34].

Applications

ANNs have been widely applied in healthcare and related fields, offering powerful tools for cancer detection, disease diagnosis, and treatment prediction. Their capacity to pick up intricate, nonlinear relationships from medical and biomedical data makes them particularly valuable in supporting medical decision-making and improving patient outcomes [79].

Beyond healthcare, ANNs are used in energy management for forecasting and optimizing consumption [80], in natural language processing for activities like sentiment analysis and translation with Transformers-like models [75], in predicting finances for market trend prediction and fraud detection [34], and in environmental engineering to enhance the efficiency of wastewater treatment processes [81].

2.1.6 Extreme Gradient Boosting (XGBoost)

XGBoost is renowned for its computational efficiency and scalability. Chen and Guestrin (2016) improved approximate tree learning's performance on big datasets by introducing a weighted quantile sketch and a unique sparsity-aware method. These developments allow XGBoost to effectively process billions of instances, even in contexts with memory constraints or distributed processing [82], [83].

XGBoost employs a collection of weak classifiers, typically DTs, to build a strong and reliable prediction model. This sequential training method incrementally merges the outputs of individual trees to build a strong model. It applies second-order gradients to reduce the loss function, enabling more accurate optimization [82].

XGBoost incorporates L1 (Lasso) and L2 (Ridge) regularization into its objective function to lessen overfitting. By limiting model complexity, these penalties encourage the use of simpler models that are more adept at generalizing to new data. In order to further avoid overfitting, it also uses a shrinkage (learning rate) parameter to control the step size during additive boosting [82].

Hyperparameter Optimization:

Effective hyperparameter tuning is crucial for maximizing XGBoost's performance. Techniques such as Bayesian optimization have been employed to optimize hyperparameters, leading to improved model accuracy and efficiency [84].

Versatility in Applications:

XGBoost's scalability and efficiency make it suitable in a variety of applications, from educational data mining to process optimization. Its capacity to manage intricate models and big datasets has been demonstrated across various domains[82].

In summary, XGBoost's design focuses on computational efficiency, robust ensemble learning, and effective regularization techniques. Its scalability and versatility have been validated across numerous studies, making it a powerful tool in the ML landscape.

2.1.7 Ensemble Modelling

Ensemble modeling is a robust ML technique that combines multiple base models to boost predictive accuracy, minimize overfitting, and enhance generalization. By aggregating diverse models, ensemble methods capitalize on the strengths of individual learners, resulting in forecasts that are more trustworthy and accurate than those produced by a single model [85].

Categories of Ensemble Models:

- **Sequential Ensembles:**

In sequential ensemble techniques, base learners are generated in sequence, with each learner attempting to correct the errors of its predecessor. This dependency among learners helps in reducing prediction errors. Notable examples include AdaBoost and Gradient Boosting[86].

- **Parallel Ensembles:**

Parallel ensemble models operate independently and simultaneously, aiming to minimize errors by averaging multiple learners. This method is exemplified by algorithms like RF, where multiple DT are constructed in parallel to improve predictive performance[87].

- **Homogeneous Ensembles:**

These ensembles consist of multiple instances of the same learning algorithm, each trained using distinct portions of the dataset or with varying parameters. A prominent example is the RF algorithm, which employs multiple DT to enhance predictive accuracy[88].

- **Heterogeneous Ensembles:**

Also known as hybrid ensembles, these merge various kinds of learning algorithms to capitalize on the advantages of each. For instance, integrating a neural network with a decision tree can enhance model diversity and performance[89].

Key Ensemble Techniques:

I. Bagging (Bootstrap Aggregating):

Bagging involves creating several variants of a predictor by generating bootstrapped samples (randomly drawn with replacement) from the original training dataset. Each model is trained independently, and their predictions are combined—usually through averaging in regression tasks or majority voting in classification—to generate the final output. This approach is designed to lower variance and mitigate overfitting. A common example is the RF algorithm[90].

II. Boosting:

Boosting is an iterative process in which each model is taught to fix the mistakes of the one before it, effectively converting weak learners into strong ones. Popular implementations include AdaBoost and Gradient Boosting, known for minimizing bias and variance, leading to improved model accuracy [91].

III. Stacking:

Stacking aggregates the predictions of several models that were trained concurrently using a meta-learner. This approach leverages the unique strengths of various algorithms, often resulting in enhanced predictive performance. However, it demands careful consideration of base and meta-learner selection to avoid overfitting[92].

Applications of Ensemble Modeling

Ensemble modeling has shown significant promise in medical and healthcare applications by enhancing the precision and reliability of predictive models. It has been used for early disease prediction, such as pneumonia, by leveraging laboratory test data to enhance diagnostic precision [93]. Additionally, ensemble techniques have been applied to predict surgical outcomes in laparoscopic procedures [94] and to improve the prediction of drug-drug interactions, contributing to safer treatment decisions [95], [96].

In other domains, ensemble modeling has enhanced sea surface temperature forecasting by integrating statistical and ML methods [97], and has improved parameter estimation in complex systems, increasing accuracy and reducing manual workload [98].

In conclusion, ensemble modeling stands as a robust approach in ML, offering enhanced predictive performance across various disciplines, including medical research and statistical analysis. Its ability to combine multiple models to mitigate individual limitations makes it a valuable tool in the development of reliable predictive systems [99].

2.2 Machine Learning Evaluation Techniques

Evaluating ML models is crucial to ensuring their predictive performance, robustness, and generalizability. Various evaluation techniques have been developed to mitigate overfitting, assess model reliability, and select the optimal algorithm for a given task. Among these, cross-validation (CV) is one of the most widely employed techniques, systematically partitioning the dataset into subsets, allowing the model to be trained on one subset and tested on another[100]. This method estimates the generalization error of the model, reducing the risk of overfitting and allows for more precise evaluation of the model's performance by providing insights into how it performs on unseen data[101][102]

2.2.1 Hold-Out Method

The Hold-Out Method is one of the simplest and most commonly used evaluation techniques in ML. It involves splitting the dataset into two distinct subsets: a training set used for model development and a test set for evaluation. The most frequently used split ratios are 70/30 or 80/20, though variations such as 60/40 or 90/10 may be employed depending on dataset size and the need for statistical reliability[100].

The primary advantage of this approach is its computational efficiency, making it well-suited for large datasets. Since the model is trained on a substantial portion of the data, it can learn meaningful patterns while the test set provides an unbiased assessment of performance[103]. However, a significant limitation is its dependency on a single random split, which can introduce high variance and bias, leading to inconsistent model evaluations. Different splits may yield different accuracy scores, making the estimation of generalization error less stable[104].

Procedure of the Hold-Out Method

I. Dataset Partitioning:

The dataset is randomly partitioned into two sets: training and test data. The test set represents previously unseen data and should be used only once to avoid biased performance estimation [100].

II. Model Selection and Training:

After data partitioning, a learning algorithm is chosen based on the problem type. The algorithm is trained on the training set, where it learns patterns and relationships among features. Hyperparameters, which influence learning behavior, must be manually specified before training, as they are not learned from the data [104].

III. Performance Evaluation on Test Set:

Once the model is trained, its generalization performance is assessed by making predictions on the test set, which acts as a proxy for real-world unseen data. The predicted class labels are then compared to the ground truth labels to compute key performance metrics such as accuracy, precision, recall, and F1-score [104].

IV. Assessing Generalization Ability:

The final step is evaluating the model's ability to generalize beyond the training data. The model's error rate on the test set provides an estimate of its expected performance on new data. However, the single-split nature of hold-out validation can lead to unreliable performance estimates when datasets are small [103].

The Hold-Out Method is preferable in scenarios where:

- I. A large dataset is available, allowing for an adequately sized test set while maintaining sufficient training data.
- II. The computational budget is limited, and cross-validation methods (which require multiple training cycles) are impractical [103].
- III. The focus is on obtaining a quick performance estimate rather than a highly precise measure of model generalization [104].

2.2.2 k-fold cross-validation

k-fold cross-validation is one of the most widely used techniques for evaluating and selecting ML models. The core idea behind cross-validation is to ensure that each data point in the dataset is used for both training and validation. In k-fold cross-validation, the dataset is partitioned into k equal subsets. During each of the k iterations, one subset serves as the validation set, while the remaining k - 1 subsets are used for training the model. This process is repeated k times, with each subset used exactly once as the validation set [103].

In this method, a learning algorithm with fixed hyperparameters is applied to train models on the training partitions in every iteration. Each fold takes a turn as the validation set, and performance metrics are recorded. After completing all k rounds, the evaluation scores are averaged to estimate the model's general performance. To determine the best set of hyperparameters, the average performance across all folds is computed for each configuration, and the one with the highest average is selected [103].

Assume the dataset presented in the equation [105] for K-folds cross validation.

$$D = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)\}$$

In this equation, x represents the characteristics, and y reflects each observation's outcome label. from 1 to n, iterations are carried out.

With this method, the k-folds CV error is illustrated in the following equation [105],:

$$K - \text{Fold CV Error} = \frac{1}{k} \sum_{i=1}^k E_i$$

Where [105],

$$E_i = \frac{1}{|D_i|} \sum_{j=1}^{|D_i|} L(y_{ij}, \widehat{y}_{ij})$$

The distinction between the "two-way" holdout approach and k-fold cross validation is that k-fold cross-validation uses every data point for both training and testing. This strategy aims to minimize pessimistic bias through the use of a larger training set instead of a big percentage of the dataset for testing[104]

For example, when $k=5$ we have 5 folds cross-validation divides the dataset into five separate subsets. This strategy involves training and validating the model several times across iterations. Each iteration uses the remaining four folds to train the model, with one chosen as the validation set. The model's performance is tested against the newly trained model. This approach validates models by using each fold only once. Five iterations are used to calculate the average model performance. In practical applications, $k=5$ is frequently chosen since it is computationally less costly than higher values of k . However, if k is too little, the model may be more sensitive to the way the data was split, adding to the variance of the estimate and increasing the pessimistic bias of the performance estimate (because there is less training data available for model fitting) [104].

2.2.3 Leave-One-Out Cross-Validation (LOOCV):

LOOCV is a type of cross-validation that uses k , where k is the number of samples in the dataset. The model is trained using the remaining dataset in each cycle, with one sample kept aside for validation [101].

Assume for this method that we have the dataset D [105], where:

$$D = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)\}$$

This strategy denotes the features with x_i and the outcome with y_{i1} for each observation i . The process iterates from 1 to n . training the model on $N-1$ number of observations, observations while leaving one out.

The outcome of each validation step is used to calculate the prediction error, typically with a predefined loss function such as Mean Squared Error (MSE) [105] Although LOOCV provides a nearly unbiased estimate of model performance, it is computationally intensive, especially for large datasets. However, it is particularly beneficial when working with small

datasets, where setting aside a validation subset would significantly reduce the amount of training data available.

2.2.4 Repeated K-Fold Cross-Validation

It is a more reliable method than regular K-Fold. It involves repeating the K-Fold procedure several times using various random partitions. This lowers variance and improves model stability, but it also raises computing costs [105]

2.2.5 Stratified Cross-Validation

A more intricate kind of cross-validation is called stratified cross-validation. Class-unbalanced datasets benefit greatly from stratified cross-validation. This guarantees that every fold preserves the class distribution of the original dataset, producing more accurate performance estimates, especially for classification tasks. There are classes in this kind of dataset [106]

These are dispersed equally over n folds, reflecting the class distribution for each fold in the original dataset. Classes may be unequally distributed in K-fold cross-validation, with certain folds having more examples than others. [101]

The k-fold stratified To ensure that each fold has (almost) the same percentage of samples from minority and majority classes as the total set, cross-validation splits the data set into k folds. The remaining folds are utilized for training, and one is selected for the test. The following steps of the process are the same as those for other cross-validation techniques [107].

Advantages of Stratified Cross-Validation

- I. **Managing Unbalanced Datasets:** Because certain classes are significantly less frequent than others, typical k-fold cross-validation may provide folds that are not reflective of the distribution as a whole in imbalanced datasets. As a result, performance metrics could lose their credibility. Using stratification guarantees that each fold's class distribution matches the original dataset's [107].
- II. **Better Assessment of Model Performance:** Because stratified k-fold cross-validation maintains the class distribution in each fold, it is possible to evaluate model performance more precisely, especially for metrics that are sensitive to class imbalance, such as accuracy, recall, and F1-score [107].
- III. **Result Reliability:** Stratification makes sure that performance evaluations don't become noticeably variable due to random fold sampling. This improves the model evaluation's consistency and dependability while reducing variability [107].

2.3 Bias-Variance Trade-off in Model Evaluation

A key idea in ML that controls model performance and generalization capacity is the bias-variance trade-off. The term "bias" describes the inaccuracy that results from oversimplifying the issue, which causes systemic errors in identifying the underlying patterns in the data. On the other hand, variance indicates a model's propensity to overfit by quantifying how sensitive it is to changes in the training data. There is an inherent tension between these two error sources: Variance frequently rises when bias is reduced, and vice versa [108]

Choosing the right ML model requires an understanding of this trade-off. While models with high variance become unduly sensitive to small changes in the training data (overfitting), models with high bias are unable to reflect the complexity of the data (underfitting). To get the best generalization performance on unseen data, the ideal model minimizes both bias and variance [108].

Mathematical Formulation of the Bias-Variance Trade-off

Bias, variance, and irreducible error are the three main components that make up the Mean Squared Error (MSE), which is frequently used to quantify the overall error of an ML model. This connection can be expressed mathematically as [108]:

$$MSE = \text{Bias}^2 + \text{Variance} + \text{IrreducibleError}$$

where:

- i. Bias: quantifies how much the expected model predictions deviate from the true function.
- ii. Variance: quantifies the degree of variation in predictions across training datasets.
- iii. Irreducible Error: represents inherent noise in the data that no model can fully eliminate.

A more detailed representation of bias squared is:

$$\text{Bias}^2 = \left(E[\hat{f}(x)] - f(x) \right)^2$$

where $E[\hat{f}(x)]$ is the expected prediction from the model across different training datasets, and $f(x)$ is the true underlying function[108].

Variance is computed as:

$$\text{Variance} = E \left[\left(\hat{f}(x) - E[\hat{f}(x)] \right)^2 \right]$$

These equations highlight that bias shows the degree to which the model's expected prediction differs from the actual function, whereas variance shows how much predictions change as a result of changes in the training data.

2.3.1 Bias-Variance Trade-off and Its Impact on Model Performance

Figure 1 illustrates how the Mean Squared Error (MSE) is influenced by the interaction between bias² and variance as model complexity increases. As the complexity of a model grows, it undergoes significant changes in its bias and variance components, directly affecting its generalization performance. Understanding this trade-off is crucial for selecting models that achieve optimal predictive accuracy without overfitting or underfitting.

At the initial phase of the graph, where model complexity is low, The model's simplicity prevents it from identifying the underlying patterns in the data. This scenario is often seen when applying a basic linear regression model to a highly nonlinear dataset. Due to its simplicity, the model exhibits high bias, meaning it makes strong assumptions about the data, causing its predictions to be systematically inaccurate. As a result, the model underfits the data, failing to learn the true relationships. The red curve in Figure 1 represents bias², which decreases as model complexity increases. This indicates that as the model becomes more flexible, it can better adapt to the training data, thereby reducing bias.

However, increasing model complexity does not always lead to better performance. As the model grows more complex, it starts capturing not only the underlying patterns in the training data but also the random noise present in the dataset. This results in higher variance, meaning the model becomes highly sensitive to small fluctuations in the training data, leading to instability in its predictions when applied to unseen data. The blue curve in Figure 1, representing variance, demonstrates an upward trend as model complexity increases, showing that highly complex models tend to overfit the training data.

The total error, represented by the black curve, forms a U-shape, illustrating the trade-off between variance and bias. Initially, when the model is too simple, high bias dominates the total error. As model complexity increases, bias decreases, but variance starts contributing more to the total error. The optimal model complexity is found at the minimum point of the U-curve, where bias and variance are optimally balanced. At this point, the model generalizes well to new data, minimizing overall prediction error.

This trade-off can be further understood through different modeling scenarios[108]:

- I. High Bias, Low Variance (Underfitting): Models with high bias make strong assumptions about the data and fail to capture its complexity. An example is linear regression applied to a highly nonlinear dataset. These models exhibit poor

performance on both training and test data due to their inability to learn from the dataset properly.

- II. Low Bias, High Variance (Overfitting): Highly sophisticated models with a lot of variance are overly sensitive to training data noise. Examples include deep decision trees and high-degree polynomial regressions. These models fit training data very well but fail to generalize to unseen data, leading to poor test performance.
- III. Optimal Trade-off: The best models lie in the middle of the U-curve, where bias² and variance are balanced, leading to minimal generalization error. At this optimal complexity, the model captures meaningful patterns in the data without being excessively sensitive to noise.

Figure 2.1 highlights why selecting an appropriate model complexity is critical in ML. It underscores the importance of regularization, cross-validation, and ensemble learning techniques to achieve a balance between bias and variance. These approaches help ensure that models generalize well to unseen data by reducing variance without introducing excessive bias.

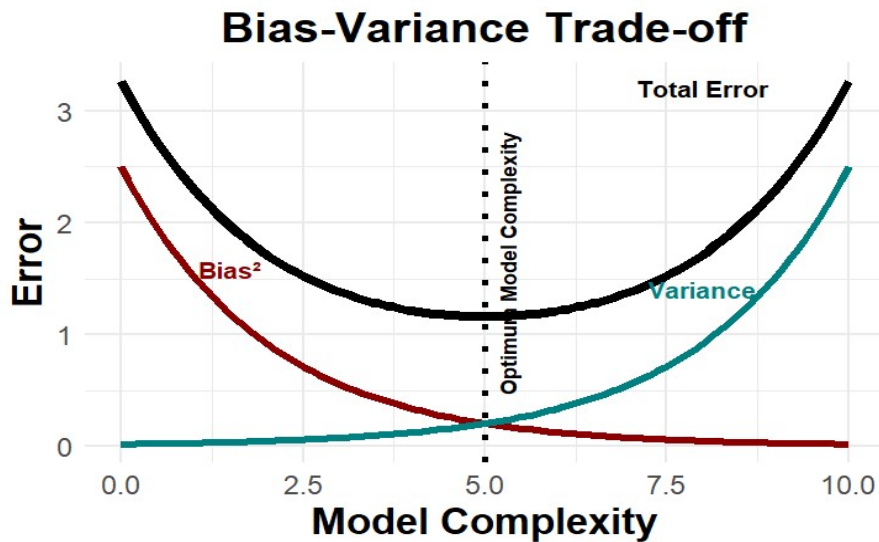


Figure 2.1 The impact of model complexity on bias, variance, and total error.

2.3.2 Bias-Variance Trade-off in Different Machine Learning Models

Empirical studies have analyzed the bias-variance decomposition across multiple algorithms. Some key findings include[108]:

- Linear Regression: Typically exhibits high bias but low variance, making it unsuitable for capturing complex patterns.

- Decision Trees: Prone to high variance, especially when deep trees are used without regularization.
- Random Forests: Reduce variance by averaging predictions from multiple decision trees, striking a better trade-off.
- Gradient Boosting (e.g., XGBoost): Reduces bias iteratively by correcting errors in weak learners, often leading to better overall performance

2.3.3 Strategies to Balance Bias and Variance

Several techniques can be used to address the bias-variance trade-off [108]:

- I. Regularization: By regulating model complexity, techniques such as Lasso (L1) and Ridge (L2) regression add penalty terms to the loss function in order to avoid overfitting.
- II. Cross-validation: Methods like k-fold cross-validation, which lower the risk of overfitting, offer more accurate estimates of test error.
- III. Ensemble Learning: Methods like Bagging (Random Forest) and Boosting (XGBoost, AdaBoost) combine multiple weak learners to achieve a better bias-variance balance trade off.
- IV. Feature Selection and Dimensionality Reduction: Eliminating irrelevant features or applying techniques like PCA reduces variance and prevents overfitting

2.4 Evaluation of Machine Learning Performance

The evaluation of ML models is a critical step in ensuring their reliability, robustness, and generalizability. Choosing the right performance measures is essential since it has a direct impact on how models are compared and understood. Based on the type of ML task, whether classification or regression, different evaluation metrics are required to measure the model's predictive ability. This section provides a comprehensive discussion on various evaluation metrics, highlighting their mathematical formulations, advantages, limitations, and real-world implications[109]

2.4.1 Classification Model Evaluation Metrics

Evaluating classification models requires a range of metrics that go beyond simple accuracy. These metrics aid in evaluating various performance-related factors, such as the ability to correctly classify minority and majority classes, handle imbalanced datasets, and provide meaningful insights into model reliability.

Confusion Matrix: Foundation of Classification Metrics

One essential technique for assessing classification models is a confusion matrix. It comprises four key elements and summarizes how the model's predictions and actual results compare:

True Positives (TP): Positive cases that were accurately predicted.

True Negatives (TN): Negative occurrences that were accurately predicted.

False Positives (FP): Instances wrongly classified as positive.

False Negatives (FN): Instances incorrectly identified as negative.

The confusion matrix provides a more comprehensive evaluation of model performance compared to a single metric such as accuracy. It serves as the foundation for deriving several classification metrics, including precision, recall, F1-score, specificity, and MCC[110]

Accuracy

Accuracy is a widely used metric for assessing classification models, representing the ratio of correctly predicted instances to the total number of observations. Mathematically, it is defined as [110]:

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

Although accuracy is a simple and widely used evaluation metric, it has notable limitations, especially when applied to imbalanced datasets [111]. It does not consider the class imbalance, allowing a model to attain high accuracy by consistently predicting the majority class while entirely overlooking the minority class [112]. This is particularly problematic in medical diagnosis scenarios, such as cancer detection, where disease prevalence is low. A model that labels all patients as “healthy” could still yield high accuracy yet be clinically ineffective by failing to identify true disease cases[113]. For instance, if only 1% of patients have a disease and the model classifies all individuals as "no disease" it would still yield 99% accuracy despite failing to identify any true cases [109]. Moreover, accuracy alone does not reflect the trade-offs between false negatives and false positives, which can have very different consequences. In applications like fraud detection or medical screening, metrics such as recall, precision, or the F1-score are often more appropriate, as they provide deeper insight into a model’s performance under such critical conditions [114], [110].

Balanced Accuracy

To mitigate the problem of imbalanced data, Balanced Accuracy is often used. This metric accounts for class imbalance by averaging the recall of each class, ensuring that both majority and minority classes contribute equally to the evaluation [113] It is defined as:

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2}$$

where sensitivity (recall) measures how well the model identifies positive cases, and specificity measures how well it identifies negative cases[115].

Precision, Recall, and F1-Score

Precision, recall, and F1-score are fundamental evaluation metrics in classification models, especially when class distributions are imbalanced. While accuracy provides an overall measure of correctness, it fails to distinguish between the costs of false positives and false negatives. Precision and recall offer more nuanced insights into model performance[113]

Precision measures the proportion of correctly predicted positive cases among all cases classified as positive:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Precision is crucial in applications where false positives have significant consequences, such as spam detection, search engines, and medical testing[112]

While Recall evaluates the model's ability to correctly identify all actual positive cases:

$$\text{Recall (Sensitivity)} = \frac{TP}{TP + FN}$$

A high recall is desirable in domains where missing a positive instance is costly, such as cancer diagnosis and fraud detection[116]

Precision and recall often exhibit an inverse relationship. Increasing precision typically lowers recall, as the model becomes more conservative in predicting positive cases, while increasing recall tends to lower precision by making the model more permissive, leading to more false positives. For example, in search engines, a high-precision model retrieves only the most relevant results but may miss useful content, whereas a high-recall model retrieves a larger set of results, including many irrelevant ones[116]. To effectively manage this trade-off, it is crucial to find a balance between precision and recall, ensuring optimal performance based on the specific application needs.

F1-score is the harmonic mean of precision and recall, balancing both metrics:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

F1-score is particularly useful in imbalanced datasets, where both false positives and false negatives are costly [113]

Precision, recall, and F1-score are applied across various domains depending on the nature of classification errors that must be minimized. In precision-oriented applications, such as spam filtering, minimizing false positives is crucial to prevent misclassification of legitimate emails[116]. Similarly, in search engines, high precision ensures that retrieved documents are relevant, improving user experience and reducing irrelevant content[116]

On the other hand, recall-oriented applications prioritize detecting all relevant instances, even at the cost of some false positives. In medical diagnosis, particularly for diseases like cancer, maximizing recall ensures that no true positive cases go undetected, potentially saving lives [109]. Similarly, in fraud detection, a high recall rate is essential to capture as many fraudulent transactions as possible, even if it results in some false positives [113]

For applications where both precision and recall are equally important, the F1-score is a more balanced metric. In text classification, achieving a balance between precision and recall ensures accurate sentiment or topic classification[109]. Likewise, in cybersecurity, maintaining an optimal F1-score helps identify security threats while minimizing false alarms [116].

The selection of the most appropriate metric depends on the cost implications of classification errors. When false positives carry significant consequences, such as in legal or financial applications, precision should be optimized. Conversely, when false negatives have a greater impact, such as in medical screening or fraud detection, recall should be prioritized. For tasks requiring an optimal trade-off between precision and recall, particularly in domains with imbalanced datasets, F1-score provides a more balanced assessment [113]

Specificity

Specificity, also known as the True Negative Rate, measures the proportion of correctly identified negative cases:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Specificity is particularly useful in situations where false positives need to be minimized. For example, in medical testing, a high specificity ensures that healthy individuals are not wrongly diagnosed with a disease, reducing unnecessary stress and treatment[116]

Matthews Correlation Coefficient (MCC)

provides a balanced measure even in imbalanced datasets:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

MCC values range from -1 (completely incorrect classification) to 1 (perfect classification). It is preferred in cases where class imbalance is severe, providing a more holistic performance measure[109]

ROC-AUC and Precision-Recall Curve

ROC-AUC measures the classifier's ability to distinguish between classes:

$$AUC = \int_0^1 TPR(FPR) d(FPR)$$

AUC values closer to 1 indicate better classification performance. The Precision-Recall (PR) curve is more informative for imbalanced datasets, where Precision is plotted against Recall. A high AUC-PR score is more indicative of effective classification than AUC-ROC when dealing with skewed class distributions[109]

2.4.2 Performance Metrics for Regression Models

Regression models predict continuous values, and their performance is evaluated using various error metrics.

Mean Absolute Error (MAE)

MAE is the average absolute difference between predicted and actual values:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

MAE provides a direct interpretation of prediction errors in the same unit as the output variable, making it easily interpretable but less sensitive to outliers[109].

Mean Squared Error (MSE) and Root Mean Squared Error (RMSE)

MSE penalizes larger errors by squaring the differences:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

RMSE takes the square root of MSE to return an error measure in the same unit as the dependent variable:

$$RMSE = \sqrt{MSE}$$

RMSE is sensitive to large errors, making it useful in applications where outliers significantly impact predictions, such as in real estate price forecasting[109].

Coefficient of Determination (R² Score)

R² measures the proportion of variance explained by the model:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Higher R² values indicate better model fit, but it can be misleading if the dataset contains non-linear relationships that are not well captured by the model[109].

3. Statistical Modeling of Colorectal Cancer Outcomes: Cancer Status, Recurrence, and Vital Status

3.1 Background

CRC is a complex and heterogeneous disease shaped by a range of clinical, demographic, and treatment-related factors. Accurately predicting key CRC outcomes—such as recurrence, current cancer status, and overall survival—is essential for guiding treatment decisions, informing follow-up strategies, and ultimately improving patient outcomes. In recent years, the application of multivariate statistical models has provided powerful tools for understanding disease trajectories and tailoring care to individual patients. This chapter presents the methodology and results of a comprehensive retrospective cohort study conducted at King Fahd Specialist Hospital in Dammam, Saudi Arabia, which aimed to predict critical CRC outcomes using both logistic regression (LR) and Cox proportional hazards models. By leveraging a well-structured and clinically rich dataset, this study contributes new insights into the prediction and management of CRC in a regional healthcare setting.

The main objective of this chapter is to develop and validate multivariate models that can accurately predict three clinically significant CRC outcomes: cancer status at last follow-up (classified as No Evidence of Disease vs. Evidence of Disease), recurrence within a six-year follow-up period, and overall survival (based on vital status: alive vs. deceased). These outcomes were selected for their central importance in cancer prognosis and patient management. LR models were used for binary classification of cancer status and recurrence, while Cox proportional hazards models were employed for time-to-event analysis of survival and recurrence-free intervals. Each model was carefully constructed and evaluated using relevant statistical performance metrics—specifically, the AUC for logistic models and Harrell’s concordance index (C-index) for survival models. Through this approach, the study aims to identify key predictors that can enhance risk stratification and inform evidence-based care strategies for CRC patients in Saudi Arabia.

The chapter begins with a description of the study design, including participant selection criteria, cohort characteristics, definitions of outcome and predictor variables, and the statistical techniques used throughout the analysis. This is followed by a presentation of the results, which include descriptive statistics of the patient population, as well as findings from multivariate LR and Cox regression models tailored to each outcome of interest. Finally, the chapter concludes with a discussion that interprets the key findings, assesses model performance and generalizability, considers clinical implications, addresses study limitations, and offers recommendations for future research.

3.2 Methods

3.2.1 Study Design and Participants

We conducted a retrospective cohort study of patients diagnosed with CRC at the King Fahd Specialist Hospital in Dammam between June 2017 and September 2024. Of the initial dataset of 1,322 patient records, 8 were excluded owing to missing values, resulting in a total of 1,314 eligible patients for analysis.

The patients were randomly partitioned into two cohorts: training cohort (n=1,052; 80%) for model development and testing cohort (n=262; 20%). This random allocation was implemented to mitigate potential biases and to ensure that both cohorts effectively represent a broader dataset.

Ethical approval for this study was obtained from the hospital's Institutional Review Board (IRB) on 03 December 2024, under reference number EXT0447. The study protocol, emphasizing strict adherence to ethical guidelines with paramount importance given to participants' well-being and privacy, was rigorously reviewed and approved by the IRB. All procedures were performed in compliance with the relevant institutional guidelines.

3.2.2 Outcome Variables

The primary outcomes of interest in this study are:

Cancer status: is a binary variable indicating whether the patient currently has active cancer at the time of last evaluation. It is classified based on medical records, diagnostic imaging, and laboratory results as either No Evidence of Disease (NED) or Evidence of Disease (EOD). NED indicates the absence of detectable cancer and is a key marker of treatment success in CRC. EOD reflects the presence of active cancer, used to monitor disease progression and treatment response in clinical studies, especially in assessments of disease-free and overall survival [117].

Recurrence: one of the primary outcomes of this study is CRC recurrence within a 6-year period. Recurrence is defined as any documented return of cancer after the patient had been treated. It is a critical metric in CRC prognosis, as it significantly impacts patient survival and informs subsequent treatment strategies. This outcome is a binary variable, categorized as 'Yes' for observed recurrence and 'No' for no recurrence during the follow-up period. [118], [119], [120].

Vital Status: Determined by patient survival status, categorized as "Alive" or "Dead". This outcome is fundamental for survival analysis and assessing the long-term impact of CRC treatments[121].

Table 3.1: Distribution of the Three Primary Outcome Variables

Outcome	Frequency	Proportion (%)
Cancer status		
No Evidence of Disease (NED)	464	35.3
Evidence of Disease (EOD)	850	64.7
Recurrence		
No recurrence	1242	94.5
Recurrence	72	5.5
Vital Status		
Alive	927	70.6
Dead	387	29.5

3.2.3 Predictor Variables

This study evaluated a comprehensive set of independent variables that may influence the three CRC outcomes: recurrence, cancer status, and survival. These variables were selected based on their clinical relevance, data availability, and established importance in prior CRC studies. They are grouped into four primary categories: demographic characteristics, tumor pathology and staging, treatment modalities, and follow-up and evaluation metrics.

- I. Demographic Variables** include age at diagnosis, sex, and marital status.
 - i. Age** is a continuous demographic variable that indicates the precise age of the patient upon diagnosis. It significantly affects the prognosis and response to treatment of CRC. According to earlier research [122], younger patients frequently have more advanced disease stages and a higher chance of recurrence. Despite this, their overall survival is often comparable to or better than older patients, emphasizing the complexity of age-related disease trajectories [122].

- ii. **Sex** is classified as male or female and has been shown to influence survival, with some studies reporting better outcomes in females [123].
 - iii. **Marital status** is classified into five categories: divorced, married, single, widowed, and unknown. This variable captures the patient's relationship status at diagnosis. Previous studies have shown that married patients, particularly married women, tend to have better survival rates [124]. This association is often attributed to stronger social support, earlier diagnosis, and greater adherence to treatment among this group. These findings highlight the importance of incorporating marital status into predictive modeling of CRC outcomes.
- II. Tumor Pathology and Staging** includes variables related to tumor biology and disease extent.
- i. **Stage at Diagnosis** is categorized as localized and in-situ, regional, distant, or unknown, reflecting disease spread at presentation. This summary stage is widely used due to its prognostic significance. Previous literature [125] has demonstrated that summary stage is one of the strongest predictors of CRC outcomes. For example, studies [125] have shown that patients diagnosed at the distant stage face a significantly elevated risk of mortality compared to those with localized disease, highlighting its critical value in outcome modeling.
 - ii. **Primary Tumor Site** is categorized into colon, rectum, rectosigmoid junction, and appendix. Each site has distinct clinical implications. Previous studies have demonstrated that rectal tumors, for example, are associated with different survival patterns compared to colon tumors[126], [127]. Although the appendix is anatomically part of the gastrointestinal tract, it is often included within the broader CRC category in research and clinical guidelines due to shared clinical features and surveillance practices. Previous Studies [128], [129]support this inclusion by jointly addressing staging and referring

to appendiceal tumors as CRC subtypes. Thus, grouping the appendix with other CRC sites aligns with established research practices and ensures consistency in data analysis.

iii. **Histology, based on ICD-O-3 coding**[130], includes adenocarcinoma NOS, mucinous adenocarcinoma, neuroendocrine and carcinoid tumors, other adenocarcinoma subtypes, and other specified or unspecified types. Adenocarcinoma NOS is the most prevalent subtype and is typically used as the reference group in analysis due to its frequency. It represents the conventional gland-forming adenocarcinoma that lacks distinctive features of other specific subtypes. Literature consistently identifies this subtype as accounting for the vast majority of CRC cases [131]. Its high prevalence and clinical heterogeneity make it a benchmark for comparison in both clinical and research settings. Literature indicates that mucinous adenocarcinoma is frequently linked to more aggressive behavior and poorer response to standard chemotherapy [132], [133]. Neuroendocrine tumors, although rare, are clinically important due to their distinct biological profile and varying survival trajectories depending on grade and stage [13].

iv. **Regional Lymph Node Involvement** is categorized as regional nodes positive, regional nodes negative, or no nodes examined. This variable serves as a key clinical staging element. Studies [134] have consistently confirmed its importance, showing that patients with node-positive disease have significantly poorer survival compared to node-negative cases, thereby reinforcing its predictive value in survival and recurrence analyses.

III. **Treatment Modalities Received** include surgical intervention, chemotherapy, radiotherapy, immunotherapy, and palliative care. Each of these variables is binary, indicating whether or not a patient received the treatment. These indicators support the evaluation of treatment effects on CRC outcomes [121]. The coding structure enables straightforward incorporation into regression models and aligns with conventions used in CRC literature [135] and large registry databases.

IV. **Follow-Up and Time-to-Event Variables** consist of date-based metrics essential for time-to-event analyses. These include the date of recurrence, last

recorded cancer status, and last contact or death. Time-to-event outcomes for the Cox regression analysis are calculated as follows:

- i. Cancer status time is calculated from the date of diagnosis to the last recorded status.
- ii. Recurrence time is computed from the date of surgery to the date of recurrence.
- iii. Vital status time is measured from the date of diagnosis to the last contact or death.

These definitions follow standard practices in survival analysis, allowing for meaningful interpretation of longitudinal data.

By comprehensively incorporating these well-structured predictors, This study seeks to determine the most relevant factors influencing CRC recurrence, cancer status, and survival in the Saudi population. This approach supports personalized care strategies and enhances evidence-based clinical decision-making.

3.2.4 Inclusion and Exclusion Criteria

Patients diagnosed with CRC and treated at King Fahd Specialist Hospital in Dammam between June 2008 and September 2024 were incorporated into this investigation. Eligible patients had confirmed diagnosis and available clinical data on key outcomes including cancer status, recurrence, and vital status.

Patients were not included if they had incomplete clinical documentation, missing critical time-to-event information, or concurrent medical conditions that could confound CRC outcomes, such as other primary cancers or severe systemic diseases. This ensured that the study outcomes were attributable specifically to CRC and not influenced by unrelated comorbidities.

3.2.5 Statistical Analysis

Frequencies and percentages were used to summarize categorical variables. Age, the only continuous variable in the dataset, was described using mean, median, and corresponding ranges. Group differences between cancer status (NED) vs. (EOD), recurrence, and overall survival (vital status) were assessed using the chi-square test for categorical variables, or Fisher's exact test when expected cell counts were less than 5. For comparisons of age between outcome groups, the independent samples *t*-test was applied, supported by the large sample size and approximate normality assumption. A *p*-value of less than 0.05 was considered statistically significant.

To evaluate predictors of recurrence, cancer status, and survival, multivariable logistic regression models were constructed for each binary outcome. All categorical predictors were converted to factors, and treatment-related variables (surgery, chemotherapy, radiation,

immunotherapy, and palliative care) were encoded as binary indicators. Reference levels for categorical variables were defined based on conventions commonly adopted in CRC literature[121], [124], [135], facilitating comparability with previous studies. For example, 'married' was used as the reference for marital status, and 'colon' for primary site. Multicollinearity was assessed using Variance Inflation Factor (VIF), and since no significant multicollinearity was observed, regularized regression models (e.g., Lasso or Ridge) were not applied.

The full dataset (N = 1,314) was randomly divided into a training cohort (80%, n = 1,052) and a testing cohort (20%, n = 262). This 80/20 split was selected to ensure an adequate volume of data for model training while preserving a separate subset for evaluating model performance on unseen data, thereby enhancing predictive accuracy and generalizability. Sequential backward feature selection (SBFS) using the Akaike Information Criterion (AIC) was employed on the training set to identify the most parsimonious subset of predictors that fit the optimized model.

For time-to-event analysis of recurrence and overall survival, multivariable Cox proportional hazards models were fitted. Recurrence time was defined as the number of days from surgery to first documented recurrence, while survival time was measured from diagnosis to death or last contact. Cancer status time was calculated as the number of days from the date of diagnosis to the date of the last recorded cancer status. Final models for all three outcomes were chosen using backward stepwise selection based on the AIC applied to the training set.

Predictive performance was thoroughly assessed for both Cox proportional hazards and LR models across all three outcomes: vital status, recurrence status, and cancer status. For the Cox models, model discrimination was evaluated using Harrell's concordance index (C-index) with 95% confidence intervals, computed on both training and test datasets. Additionally, recurrence-free survival curves were generated by comparing the actual Kaplan–Meier survival estimates from the training set with the predicted survival probabilities from the test set, using representative patient profiles. This allowed for a visual inspection of calibration between observed and predicted survival trends.

For the LR models, predictive accuracy was evaluated using the AUC along with its 95% confidence intervals. ROC curves were plotted for both training and test sets across all three outcomes. These evaluations enabled consistent and comparative assessment of model performance between development and validation cohorts.

All statistical analyses were conducted using R (version 4.4.2). Key packages included dplyr, readxl, and stringr for data handling; caret, MASS, and car for model development and feature selection; survival and survminer for survival analysis; and pROC and PRROC for model evaluation. The writexl package was used to export results.

Table 3.2: Description of the variables in the dataset.

Variable	Definition	Value
Gender	Refers to Patient Gender	
Male		715 (54.4%)
Female		599 (45.6%)
Marital status	Refers to Marital Status at Diagnosis	
Divorced		24 (1.8%)
Married		1080 (82.2%)
Single		104 (7.9%)
Widowed		37 (2.8%)
Unknown		69 (5.3%)
Regional Lymph Node Involvement	Status of regional lymph nodes, based on surgical removal, biopsy, or medical records — includes both examined and unexamined nodes, and whether any were found positive.	
No Nodes Examined		469 (35.7%)
Regional Nodes Negative		432 (32.9%)
Regional Nodes Positive		413 (31.4%)
Stage at Diagnosis	Refers to to a system used to describe the extent of cancer spread from its point of origin.	
Distant		427 (32.5%)
Localized and In-Situ		207 (15.8%)
Regional		674 (51.3%)
Unknown		6 (0.5)

Primary Tumor Site	Refers to the anatomical location of the primary tumor.	
Colon		723 (55.0%)
Rectosigmoid		178 (13.6%)
Rectum		389 (29.6%)
Appendix		24 (1.8%)
Histology – Behavior	Refers to the classification of tumors based on their microscopic composition and their biological behavior	
Adenocarcinoma NOS		1180 (89.8%)
Mucinous Adenocarcinoma		62 (4.7%)
Neuroendocrine & Carcinoid Tumors		27 (2.1%)
Other Adenocarcinoma Subtypes		26 (2.0%)
Others		19 (1.5%)
Age	Refers to Age at Diagnosis in years at the time of disease diagnosis	
Median		56
Mean		55.8
Range		9 – 100
Surgery	This specifies whether the patient received surgical treatment.	
No		376 (28.6%)
Yes		938 (71.4%)
Radiation	This checks whether the patient received radiation treatment.	
No		1093 (83.2%)

Yes		221 (16.8%)
Chemotherapy		
This determines whether the patient has had chemotherapy.		
No		359 (27.3%)
Yes		955 (72.7%)
Immunotherapy		
This determines if the patient has had immunotherapy or not.		
No		1276 (97.1%)
Yes		38 (2.9%)
Palliative		
Palliative care focuses on enhancing quality of life by alleviating symptoms such as pain or obstruction, particularly in advanced or metastatic cancer cases. This variable indicates whether the patient received palliative care or not.		
No		1208 (91.9%)
Yes		106 (8.1%)

3.3 Result

3.3.1 Descriptive Characteristics of the Cohort

The study cohort comprised 1,314 patients diagnosed with CRC. Table 3.1 summarizes the distribution of the three primary outcomes of interest. At the end of follow-up, 29.5% of patients had died ($n = 387$), while 70.6% were alive ($n = 927$). Regarding cancer status, 64.7% ($n = 850$) had evidence of disease (EOD), whereas 35.3% ($n = 464$) were classified as having no evidence of disease (NED). Recurrence was documented in 5.5% of the cohort ($n = 72$), with the remaining 94.5% ($n = 1,242$) exhibiting no recurrence.

We next examined baseline characteristics stratified by each of the three key clinical outcomes: cancer status, recurrence, and vital status.

I. Patient Characteristics by Cancer Status

Among the 1,314 CRC patients, 464 (35.3%) were classified as having No Evidence of Disease (NED), while 850 (64.7%) were recognized as having Evidence of Disease (EOD) at their most recent follow-up. Table 3.3 presents the distribution of demographic and clinical profile across these two cancer status categories and highlights the significant differences observed.

Patients with EOD had a significantly higher mean age (56.6 years) in contrast to those in the NED group (54.4 years) ($p = 0.0053$). Gender distribution was similar between the groups, with males comprising slightly more than half in both NED (53.9%) and EOD (54.7%) categories ($p = 0.8184$). Likewise, marital status distributions were not significantly different ($p = 0.9599$), with married individuals forming the predominant group in both cancer status categories.

Regarding clinical variables, lymph node involvement was significantly related to cancer status ($p < 0.001$). Notably, nearly half (48.4%) of EOD patients had no lymph nodes examined, compared to only 12.5% of NED patients. Conversely, regional node-negative findings more commonly observed in the NED group (53.9%) than the EOD group (21.4%), while regional node-positive status was relatively similar.

Stage at diagnosis also differed markedly between the groups ($p < 0.001$). The majority of NED patients were diagnosed at the regional stage (67.0%), whereas the EOD group had a substantially higher proportion of distant stage cases (46.7%) and a lower proportion of localized stage cases

Table 3.3: Comparison of patients' characteristics between the No Evidence of Disease (NED) And Evidence of Disease (EOD) cases

Predictors	NED	EOD	Total	P-value
Age				0.005331
Mean	54.42026	56.57765		
(Range)	(9 – 88)	(10 – 100)		
Gender				0.8184
Male	250 (53.9%)	465 (54.7%)	715 (54.4%)	
Female	214 (46.1%)	385 (45.3%)	599 (45.6%)	
Marital status				0.9599
Divorced	8 (1.7%)	16 (1.9%)	24 (1.8%)	
Married	382 (82.3%)	698 (82.1%)	1080 (82.2%)	
Single	34 (7.3%)	70 (8.2%)	104 (7.9%)	
Widowed	14 (3.0%)	23 (2.7%)	37 (2.8%)	
Unknown	26 (5.6%)	43 (5.1%)	69 (5.3%)	
Regional Lymph Node Involvement				0
No Nodes Examined	58 (12.5%)	411 (48.4%)	469 (35.7%)	
Regional Nodes Negative	250 (53.9%)	182 (21.4%)	432 (32.9%)	
Regional Nodes Positive	156 (33.6%)	257 (30.2%)	413 (31.4%)	
Stage at Diagnosis				0.000
Distant	30 (6.5%)	397 (46.7%)	427 (32.5%)	
Localized and In-Situ	123 (26.5%)	84 (9.9%)	207 (15.8%)	
Regional	311 (67.0%)	363 (42.7%)	674 (51.3%)	
Unknown	0 (0.0%)	6 (0.7%)	6 (0.5%)	
Primary Tumor Site				0.0033

Colon	272 (58.6%)	451 (53.1%)	723 (55.0%)	
Rectosigmoid	73 (15.7%)	105 (12.4%)	178 (13.5%)	
Rectum	109 (23.5%)	280 (32.9%)	389 (29.6%)	
Appendix	10 (2.2%)	14 (1.6%)	24 (1.8%)	
Histology Behavior				0.5144
Adenocarcinoma NOS	420 (90.5%)	760 (89.4%)	1180 (89.8%)	
Mucinous Adenocarcinoma	22 (4.7%)	40 (4.7%)	62 (4.7%)	
Neuroendocrine & Carcinoid Tumors	10 (2.2%)	17 (2.0%)	27 (2.1%)	
Other Adenocarcinoma Subtypes	9 (1.9%)	17 (2.0%)	26 (2.0%)	
Others	3 (0.6%)	16 (1.9%)	19 (1.4%)	
Surgery				0.000
No	13 (2.8%)	363 (42.7%)	376 (28.6%)	
Yes	451 (97.2%)	487 (57.3%)	938 (71.4%)	
Radiation				0.077
No	374 (80.6%)	719 (84.6%)	1093 (83.2%)	
Yes	90 (19.4%)	131 (15.4%)	221 (16.8%)	
Chemotherapy				0.458
No	133 (28.7%)	226 (26.6%)	359 (27.3%)	
Yes	331 (71.3%)	624 (73.4%)	955 (72.7%)	
Immunotherapy				0.0172
No	458 (98.7%)	818 (96.2%)	1276 (97.1%)	
Yes	6 (1.3%)	32 (3.8%)	38 (2.9%)	
Palliative				0.000
No	462 (99.6%)	746 (87.8%)	1208 (91.9%)	

Yes	2 (0.4%)	104 (12.2%)	106 (8.1%)
-----	----------	-------------	------------

(9.9%). The NED group had a significantly greater representation of localized and in-situ diagnoses (26.5%) compared to EOD patients (9.9%).

The primary tumor site showed a significant relationship with cancer status ($p = 0.0033$). Rectal tumors were more common among EOD cases (32.9%) compared to NED (23.5%), while colon tumors were more frequent in the NED group (58.6%) than in EOD (53.1%). Appendix and rectosigmoid sites were relatively rare in both groups.

Although histologic subtype distributions were not significantly different ($p = 0.5144$), Adenocarcinoma NOS remained the most common histological type in both groups, followed by mucinous adenocarcinoma. More aggressive or rare histologies (e.g., “others”) were slightly more frequent in the EOD group.

Surgical treatment exhibited a strong association with cancer status ($p < 0.001$), as nearly all NED patients underwent surgery (97.2%) compared to just 57.3% of EOD patients. Radiation therapy did not differ significantly between groups ($p = 0.077$), while chemotherapy usage was consistent across NED (71.3%) and EOD (73.4%) groups ($p = 0.458$). Immunotherapy was significantly more common in the EOD group (3.8%) than in the NED group (1.3%) ($p = 0.0172$).

Palliative care was notably more frequent in the EOD group, with 12.2% receiving such care, compared to just 0.4% in the NED group ($p < 0.001$), highlighting its strong association with ongoing disease burden and advanced clinical status.

II. Patient Characteristics by Recurrence Status

Among the 1,314 CRC patients analyzed, 72 (5.5%) experienced recurrence during the follow-up period, with a median recurrence time of 604 days (approximately 1.7 years). The majority of recurrence events occurred within three years of treatment. Table 3.4 presents the comparison of baseline characteristics between recurrent and non-recurrent patients.

Patients who developed recurrence were younger on average (mean age = 51.0 years) than those without recurrence (mean age = 56.1 years), and this difference was statistically significant ($p = 0.001$). No significant difference in sex distribution was observed between the two groups ($p = 0.2549$), with males comprising slightly more than half of the cohort in both recurrence and non-recurrence groups. regarding marital status, most patients in both groups were married, and the distribution did not differ significantly ($p = 0.823$).

Regarding clinical characteristics, a significant difference was observed in lymph node involvement ($p = 0.0009$). Patients with recurrence were more likely to have positive lymph nodes

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

Predictors	Non-recurrence	Recurrence	Total	P-value
Age				0.001
Mean	56.1	51.0		
(Range)	(9 – 100)	(19 – 80)		
Gender				0.2549
Male	681 (54.8%)	34 (47.2%)	715 (54.4%)	
Female	561 (45.2%)	38 (52.8%)	599 (45.6%)	
Marital status				0.823
Divorced	23 (1.9%)	1 (1.4%)	24 (1.8%)	
Married	1017 (81.9%)	63 (87.5%)	1080 (82.2%)	
Single	101 (8.1%)	3 (4.2%)	104 (7.9%)	
Widowed	35 (2.8%)	2 (2.8%)	37 (2.8%)	
Unknown	66 (5.3%)	3 (4.2%)	69 (5.3%)	
Regional Lymph Node Involvement				0.0009
No Nodes Examined	458 (36.9%)	11 (15.3%)	469 (35.7%)	
Regional Nodes Negative	402 (32.4%)	30 (41.7%)	432 (32.9%)	
Regional Nodes Positive	382 (30.8%)	31 (43.1%)	413 (31.4%)	
Stage at Diagnosis				
Distant	418 (33.7%)	9 (12.5%)	427 (32.5%)	0.000
Localized and In-Situ	197 (15.9%)	10 (13.9%)	207 (15.8%)	

Regional	624 (50.2%)	50 (69.4%)	674 (51.3%)	
Unknown	3 (0.2%)	3 (4.2%)	6 (0.5%)	
Primary Tumor Site				0.2217
Colon	676 (54.4%)	47 (65.3%)	723 (55.0%)	
Rectosigmoid	168 (13.5%)	10 (13.9%)	178 (13.5%)	
Rectum	374 (30.1%)	15 (20.8%)	389 (29.6%)	
Appendix	24 (1.9%)	0 (0.0%)	24 (1.8%)	
Histology Behavior				0.0836
Adenocarcinoma NOS	1119 (90.1%)	61 (84.7%)	1180 (89.8%)	
Mucinous Adenocarcinoma	55 (4.4%)	7 (9.7%)	62 (4.7%)	
Neuroendocrine & Carcinoid Tumors	27 (2.2%)	0 (0.0%)	27 (2.1%)	
Other Adenocarcinoma Subtypes	23 (1.9%)	3 (4.2%)	26 (2.0%)	
Others	18 (1.4%)	1 (1.4%)	19 (1.4%)	
Surgery				0.000
No	374 (30.1%)	2 (2.8%)	376 (28.6%)	
Yes	868 (69.9%)	70 (97.2%)	938 (71.4%)	
Radiation				0.3977
No	1030 (82.9%)	63 (87.5%)	1093 (83.2%)	
Yes	212 (17.1%)	9 (12.5%)	221 (16.8%)	
Chemotherapy				0.000
No	355 (28.6%)	4 (5.6%)	359 (27.3%)	
Yes	887 (71.4%)	68 (94.4%)	955 (72.7%)	
Immunotherapy				1

No	1206 (97.1%)	70 (97.2%)	1276 (97.1%)
Yes	36 (2.9%)	2 (2.8%)	38 (2.9%)
Palliative			0.8909
No	1141 (91.9%)	67 (93.1%)	1208 (91.9%)
Yes	101 (8.1%)	5 (6.9%)	106 (8.1%)

(43.1%) compared to those without recurrence (30.8%). Additionally, the stage at diagnosis exhibited a strong association with recurrence ($p < 0.001$). While regional stage was the most common in both groups, it was notably more prevalent among those who recurred (69.4% vs. 50.2%). Conversely, distant stage disease was less common in the recurrence group (12.5%) than in non-recurrence cases (33.7%), potentially reflecting stage-based treatment dynamics and follow-up intensity.

No statistically significant differences were detected with respect to primary tumor site ($p = 0.222$), though recurrence cases were slightly more concentrated in the colon. Histological subtype showed a borderline association ($p = 0.084$), with a higher proportion of mucinous adenocarcinomas observed in recurrence cases (9.7% vs. 4.4%). Notably, surgical treatment was significantly more frequent among patients with recurrence (97.2% vs. 69.9%; $p < 0.001$), as was chemotherapy (94.4% vs. 71.4%; $p < 0.001$), possibly reflecting clinical prioritization for aggressive treatment in patients at higher risk. Radiation therapy and immunotherapy usage did not differ significantly between groups, nor did the use of palliative care services.

Together, these findings highlight significant differences in age, lymph node status, disease stage, and treatment patterns between recurrence and non-recurrence cases, offering insights into clinical profiles that may be associated with elevated recurrence risk.

III. Patient Characteristics by Vital Status

Of the total cohort of 1,314 CRC patients, 387 (29.5%) were deceased at last follow-up, while 927 (70.5%) were alive. Table 3.5 outlines the baseline characteristics of these two groups and highlights statistically significant differences across several clinical and demographic variables.

The mean age of deceased patients was slightly higher (57.1 years) than that of those alive (55.3 years), with this difference reaching statistical significance ($p = 0.0328$). Sex distribution was comparable between the groups, with males representing the majority in both categories ($p = 0.5369$). Marital status distributions were not significantly different (p

= 0.1691), with married individuals comprising the largest group across both survival outcomes.

Regarding clinical characteristics, a notable difference was observed in lymph node evaluation and involvement ($p < 0.001$). Specifically, 64.6% of deceased patients had no lymph nodes examined, compared to only 23.6% of living patients. Conversely, the living group had a greater proportion of negative node findings (42.2% vs. 10.6%).

Table 3.5: Comparison of patients' characteristics between two categories of Vital Status "Alive" and "Dead" Cases

Predictors	Dead	Alive	Total	P-value
Age				0.03279
Mean	57.1	55.3		
(Range)	(19 – 95)	(9 – 100)		
Gender				0.5369
Male	205 (53%)	510 (55%)	715 (54.4%)	
Female	182 (47%)	417 (45%)	599 (45.6%)	
Marital status				0.1691
Divorced	9 (2.3%)	15 (1.6%)	24 (1.8%)	
Married	324 (83.7%)	756 (81.6%)	1080 (82.2%)	
Single	29 (7.5%)	75 (8.1%)	104 (7.9%)	
Widowed	13 (3.4%)	24 (2.6%)	37 (2.8%)	
Unknown	12 (3.1%)	57 (6.1%)	69 (5.3%)	
Regional Lymph Node Involvement				0.000
No Nodes Examined	250 (64.6%)	219 (23.6%)	469 (35.7%)	
Regional Nodes Negative	41 (10.6%)	391 (42.2%)	432 (32.9%)	
Regional Nodes Positive	96 (24.8%)	317 (34.2%)	413 (31.4%)	
Stage at Diagnosis				0.000

Distant	268 (69.3%)	159 (17.2%)	427 (32.5%)	
Localized and In-Situ	18 (4.7%)	189 (20.4%)	207 (15.8%)	
Regional	97 (25.1%)	577 (62.2%)	674 (51.3%)	
Unknown	4 (1.0%)	2 (0.2%)	6 (0.5%)	
Primary Tumor Site				0.0462
Colon	228 (58.9%)	495 (53.4%)	723 (55.0%)	
Rectosigmoid	48 (12.4%)	130 (14.0%)	178 (13.5%)	
Rectum	109 (28.2%)	280 (30.2%)	389 (29.6%)	
Appendix	2 (0.5%)	22 (2.4%)	24 (1.8%)	
Histology Behavior				0.0003
Adenocarcinoma NOS	330 (85.3%)	850 (91.7%)	1180 (89.8%)	
Mucinous Adenocarcinoma	24 (6.2%)	38 (4.1%)	62 (4.7%)	
Neuroendocrine & Carcinoid Tumors	8 (2.1%)	19 (2.0%)	27 (2.1%)	
Other Adenocarcinoma Subtypes	12 (3.1%)	14 (1.5%)	26 (2.0%)	
Others	13 (3.4%)	6 (0.6%)	19 (1.4%)	
Surgery				0.000
No	249 (64.3%)	127 (13.7%)	376 (28.6%)	
Yes	138 (35.7%)	800 (86.3%)	938 (71.4%)	
Radiation				0.0015
No	342 (88.4%)	751 (81%)	1093 (83.2%)	
Yes	45 (11.6%)	176 (19%)	221 (16.8%)	
Chemotherapy				0.867
No	104 (26.9%)	255 (27.5%)	359 (27.3%)	
Yes	283 (73.1%)	672 (72.5%)	955 (72.7%)	

Immunotherapy			0.0552
No	370 (95.6%)	906 (97.7%)	1276 (97.1%)
Yes	17 (4.4%)	21 (2.3%)	38 (2.9%)
Palliative			0.000
No	305 (78.8%)	903 (97.4%)	1208 (91.9%)
Yes	82 (21.2%)	24 (2.6%)	106 (8.1%)

Stage at diagnosis was strongly associated with vital status ($p < 0.001$). Most patients who had died were diagnosed at the distant stage (69.3%), whereas the majority of those alive were diagnosed at the regional stage (62.2%). Localized and in-situ cases were more prevalent among living patients (20.4%) than deceased (4.7%).

Tumor site also showed a statistically significant difference ($p = 0.0462$), with appendix tumors being slightly more common in the living group (2.4%) compared to the deceased (0.5%). Although colon remained the most frequent tumor site in both groups, rectal tumors were also prevalent.

Histological behavior showed a statistically significant association with survival ($p = 0.0003$). Adenocarcinoma NOS remained the dominant subtype in both groups, though it was more common among living patients (91.7%) compared to deceased (85.3%). Less common and more aggressive subtypes (e.g., mucinous adenocarcinoma and “others”) were slightly more frequent in the deceased group.

Surgical treatment was significantly less common in deceased patients (35.7%) than in those alive (86.3%) ($p < 0.001$). Similarly, the absence of radiation therapy was more frequent among deceased patients ($p = 0.0015$), whereas chemotherapy use did not differ significantly between the groups ($p = 0.867$). Immunotherapy use approached statistical significance ($p = 0.0552$), with a higher proportion of deceased patients receiving it. Palliative care was substantially more common among the deceased (21.2%) compared to living patients (2.6%) ($p < 0.001$), reflecting its role in end-of-life care.

3.3.2 Multivariate Logistic Regression Analysis for Cancer Status

After variable selection, the final multivariate logistic regression model for predicting cancer status (EOD) vs. (NED) included age, site category, stage at diagnosis, regional lymph node involvement, surgery, radiation therapy, and palliative care as predictors (Table 3.6). The model showed strong discriminative performance, achieving an AUC of 0.81 (95% CI: 0.79–0.84) in the training set and 0.81 (95% CI: 0.76–0.86) in the test set (Figure 3.1).

Older age was marginally associated with higher odds of having EOD (AOR = 1.01; 95% CI: 1.00–1.02; $p = 0.075$). Tumor site also emerged as a significant predictor: patients with tumors in the rectum had increased odds of having EOD compared to those with colon tumors (AOR = 2.09; 95% CI: 1.32–3.36; $p = 0.002$). Stage at diagnosis was a strong and consistent predictor—patients diagnosed at the distant stage were substantially more likely to exhibit EOD (AOR = 7.72; 95% CI: 4.36–14.01; $p < 0.001$), while those at the regional stage also had elevated odds, though marginally significant (AOR = 1.54; 95% CI: 1.00–2.39; $p = 0.052$).

Table 3.6: The result of the multivariate logistic regression model for cancer status with the adjusted odd ratio with 95% CI.

Predictors	β (SE)	AOR	95% CI	p-value
Age	0.010 (0.006)	1.01	0.999–1.022	0.075
Site Category				
Colon	Ref	1		
APPENDIX	0.132 (0.577)	1.14	0.365–3.596	0.820
RECTOSIGMOID	-0.230 (0.232)	0.79	0.503–1.248	0.320
RECTUM	0.738 (0.238)	2.09	1.318–3.361	0.002
Stage Category				
Localized and In-Situ	Ref	1		
Regional	0.433 (0.223)	1.54	0.998–2.393	0.052
Distant	2.044 (0.297)	7.72	4.358–14.007	< 0.001
Unknown	14.816 (378.455)	2,701,173	2.64×10^{-316} – 1.09×10^{328}	0.969
Regional Nodes				
Regional Nodes Negative	Ref	1		
No Nodes Examined	0.704 (0.260)	2.02	1.216–3.381	0.007
Regional Nodes Positive	0.419 (0.191)	1.52	1.048–2.212	0.028

Surgery				
No	Ref	1		
Yes	-1.893 (0.383)	0.15	0.068–0.309	< 0.001
Radiation				
No	Ref	1		
Yes	-0.790 (0.259)	0.45	0.272–0.751	0.002
Palliative				
No	Ref	1		
Yes	1.636 (0.775)	5.13	1.377–33.562	0.035

Lymph node evaluation revealed important associations. Patients with no lymph nodes examined had significantly higher odds of being classified as EOD (AOR = 2.02; 95% CI: 1.22–3.38; $p = 0.007$), as did those with regional node positivity (AOR = 1.52; 95% CI: 1.05–2.21; $p = 0.028$), compared to those with negative lymph nodes.

Treatment-related variables also contributed meaningfully. Surgery was strongly protective, with those undergoing surgery having significantly lower odds of EOD (AOR = 0.15; 95% CI: 0.07–0.31; $p < 0.001$). Similarly, radiation therapy was associated with reduced odds of EOD (AOR = 0.45; 95% CI: 0.27–0.75; $p = 0.002$). On the other hand, palliative care was positively associated with the presence of disease (AOR = 5.13; 95% CI: 1.38–33.56; $p = 0.035$), consistent with its role in managing patients with advanced or persistent cancer.

In summary, the model highlighted several independent predictors of cancer status, including tumor site, disease stage, lymph node involvement, and treatment history. These findings may support improved prognostic stratification and inform post-treatment surveillance strategies in CRC care.

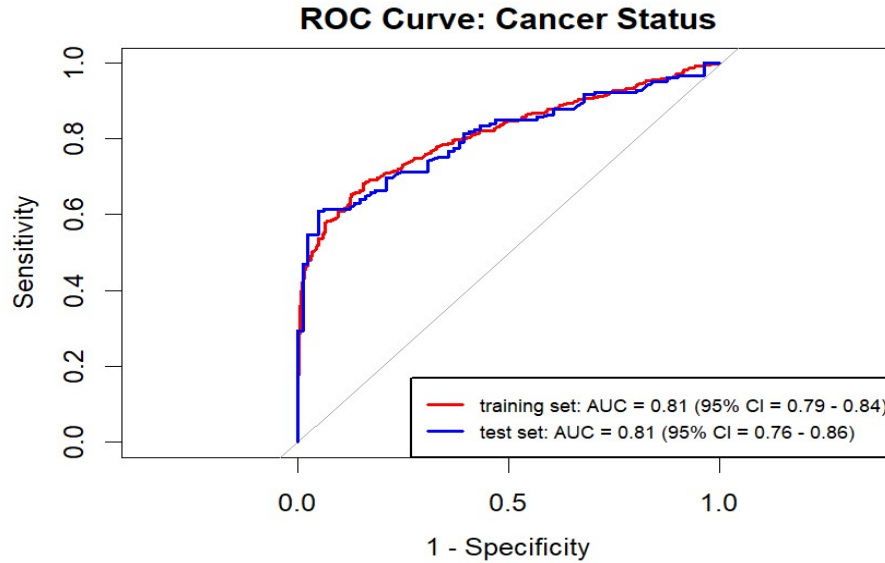


Figure 3.1: ROC curve of the logistic regression model showing the AUC values of the training and testing set for Cancer Status.

3.3.3 Multivariate Logistic Regression Analysis for Recurrence Status

The final LR model for recurrence included age, stage at diagnosis, surgery, radiation therapy, and chemotherapy as informative predictors (Table 3.7). The model demonstrated good discriminative ability, with an AUC of 0.80 (95% CI: 0.75–0.85) for the training set and 0.71 (95% CI: 0.59–0.83) for the test set (Figure 3.2).

Table 3.7: The result of the multivariate logistic regression model for recurrence with the adjusted odd ratio with 95% CI.

Predictors	β (SE)	AOR	95% CI	p-value
Age	-0.034 (0.011)	0.97	0.945–0.988	0.003
Stage Category				
Localized and In-Situ	Ref	1		
Regional	-0.057 (0.456)	0.94	0.407–2.495	0.901
Distant	-0.933 (0.593)	0.39	0.120–1.280	0.116
Unknown	3.244 (1.067)	25.65	3.293–258.329	0.002
Surgery				

No	Ref	1		
Yes	1.939 (0.761)	6.95	1.935–44.678	0.011
Radiation				
No	Ref	1		
Yes	-0.679 (0.408)	0.51	0.212–1.069	0.096
Chemotherapy				
No	Ref	1		
Yes	2.491 (0.700)	12.07	3.606–60.846	< 0.001

Age was significantly associated with recurrence status (AOR = 0.97, 95% CI: 0.945–0.988, $p = 0.003$), suggesting that the odds of recurrence decreased by approximately 3% with each additional year of age. Patients with unknown cancer stage had significantly elevated odds of recurrence compared to those with localized/in-situ disease (AOR = 25.65; 95% CI: 3.29–258.33; $p = 0.002$). While the regional and distant stages did not reach statistical significance, distant-stage disease was associated with lower odds of recurrence (AOR = 0.39; $p = 0.116$), potentially due to more aggressive initial management or loss to follow-up.

Treatment variables were also influential. Undergoing surgery was significantly associated with higher odds of documented recurrence (AOR = 6.95; 95% CI: 1.94–44.68; $p = 0.011$), potentially reflecting that patients receiving surgical intervention were those more likely to be followed closely and thus have recurrence detected. Similarly, chemotherapy was strongly associated with recurrence (AOR = 12.07; 95% CI: 3.61–60.85; $p < 0.001$), which may again be reflective of treatment intensity in higher-risk patients. Radiation therapy showed a borderline association with reduced recurrence risk (AOR = 0.51; 95% CI: 0.21–1.07; $p = 0.096$).

In summary, the model identified age, cancer stage, surgical intervention, and chemotherapy as significant predictors of recurrence status, highlighting clinical characteristics and treatment patterns that may help inform risk stratification.

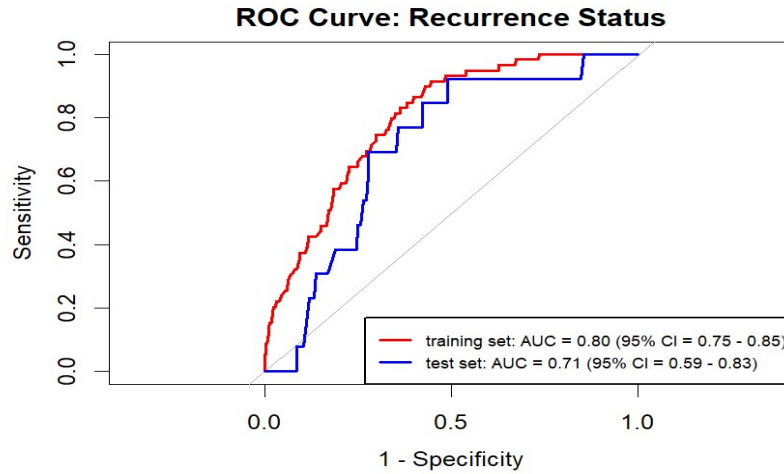


Figure 3.2: ROC curve of the logistic regression model showing the AUC values of the training and testing set for Recurrence Status.

3.3.4 Multivariate Logistic Regression Analysis for Vital Status

The final LR model for vital status included histological subtype, tumor site, cancer stage, Regional Lymph Node Involvement, surgery, immunotherapy, and palliative care as informative predictors (Table 3.8). The model exhibited strong discriminative performance, with an AUC of 0.85 (95% CI: 0.83–0.88) for the training set and 0.84 (95% CI: 0.79–0.90) for the test set (Figure 3.3).

Table 3.8: The result of the multivariate logistic regression model for vital with the adjusted odd ratio with 95% CI.

Predictors	β (SE)	AOR	95% CI	p-value
Histology Category				
Adenocarcinoma NOS	Ref	1		
Mucinous Adenocarcinoma	1.201 (0.355)	3.32	1.646–6.656	0.001
Neuroendocrine & Carcinoid Tumors	0.0004 (0.635)	1.00	0.285–3.485	0.999
Other Adenocarcinoma Subtypes	0.721 (0.542)	2.06	0.686–5.847	0.184
Others	1.024 (0.685)	2.78	0.758–11.579	0.135
Site Category				
Colon	Ref	1		
APPENDIX	-1.850 (1.006)	0.16	0.016–0.925	0.066
RECTOSIGMOID	-0.604 (0.265)	0.55	0.322–0.911	0.023
RECTUM	-0.500 (0.210)	0.61	0.400–0.911	0.017
Stage Category				
Localized and In-Situ	Ref	1		
Regional	0.431 (0.348)	1.54	0.794–3.131	0.216
Distant	1.801 (0.351)	6.06	3.107–12.405	< 0.0001
Unknown	3.195 (1.197)	24.40	2.994–519.519	0.008
Regional Nodes				
Regional Nodes Negative	Ref	1		
No Nodes Examined	0.687 (0.320)	1.99	1.054–3.712	0.032
Regional Nodes Positive	0.584 (0.259)	1.79	1.089–3.011	0.024
Surgery				
No	Ref	1		

Yes	-1.549 (0.277)	0.21	0.122–0.362	< 0.0001
Immunotherapy				
No	Ref	1		
Yes	-0.711 (0.442)	0.49	0.205–1.172	0.108
Palliative				
No	Ref	1		
Yes	1.009 (0.322)	2.74	1.483–5.272	0.002

Several predictors demonstrated significant associations with survival. Patients with mucinous adenocarcinoma had more than threefold increased odds of death compared to those with adenocarcinoma NOS (AOR = 3.32; 95% CI: 1.65–6.66; $p = 0.001$), suggesting a more aggressive disease course. Tumor site also influenced mortality, with rectum (AOR = 0.61; $p = 0.017$) and rectosigmoid (AOR = 0.55; $p = 0.023$) tumors associated with significantly lower odds of death compared to colon cancer. Although appendix tumors showed the lowest odds (AOR = 0.16), this was only marginally significant ($p = 0.066$).

Stage at diagnosis was a strong predictor of vital status. Patients diagnosed at the distant stage had sixfold higher odds of death compared to those with localized/in-situ disease (AOR = 6.06; 95% CI: 3.11–12.41; $p < 0.001$), while those with unknown stage had dramatically elevated odds (AOR = 24.40; 95% CI: 2.99–519.52; $p = 0.008$). Lymph node involvement was also informative; patients with no nodes examined (AOR = 1.99; $p = 0.032$) or positive regional nodes (AOR = 1.79; $p = 0.024$) were at increased risk of death compared to those with negative nodes.

Regarding treatment, surgery was strongly associated with reduced mortality. Patients who underwent surgical resection had significantly lower odds of death (AOR = 0.21; 95% CI: 0.12–0.36; $p < 0.001$). Palliative care was significantly associated with higher odds of death (AOR = 2.74; 95% CI: 1.48–5.27; $p = 0.002$), consistent with its role in advanced disease management. Immunotherapy showed a borderline protective effect but did not reach statistical significance (AOR = 0.49; $p = 0.108$).

In summary, this model highlighted key clinical predictors of mortality in CRC patients, including tumor histology, stage at diagnosis, nodal involvement, surgical intervention, and palliative care utilization. These findings underscore the prognostic value of initial disease presentation and treatment patterns in shaping survival outcomes.

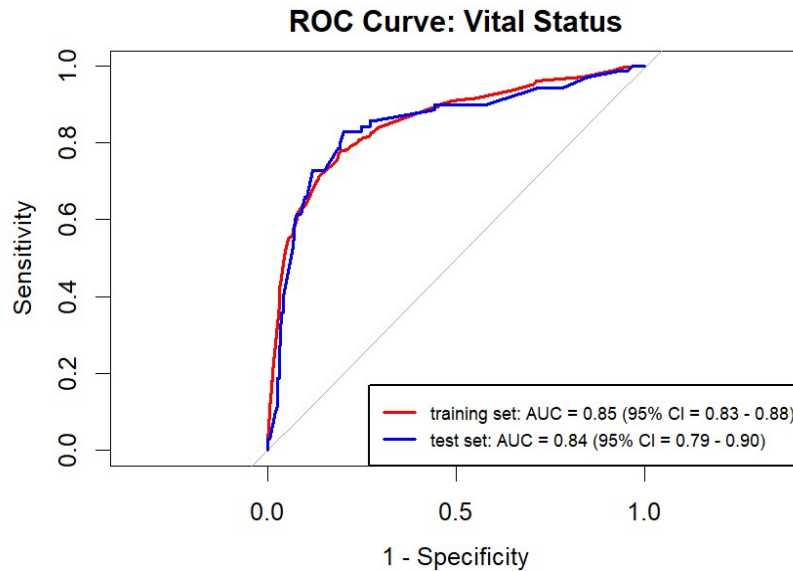


Figure 3.3: ROC curve of the logistic regression model showing the AUC values of the training and testing set for Vital Status.

3.3.5 Multivariable Cox Proportional Hazards Analysis for cancer status

This section presents the results of the multivariable Cox proportional hazards model assessing predictors of cancer status (Table 3.9), where the event is defined as Evidence of Disease (EOD). The model demonstrated solid discriminatory performance, achieving a C-index of 0.73 (95% CI: 0.71–0.75) in the training set and 0.68 (95% CI: 0.64–0.73) in the test set, indicating acceptable generalizability.

Age was a statistically significant predictor of persistent disease. A one-year increase in age was associated with a 1% increase in the hazard of being classified as EOD (HR = 1.01; 95% CI: 1.002–1.015; $p = 0.007$).

In terms of histological subtype, patients with Neuroendocrine and Carcinoid tumors had a significantly lower hazard of EOD compared to those with adenocarcinoma NOS (HR = 0.56; 95% CI: 0.32–0.96; $p = 0.034$). The “Others” subtype also showed a strong protective association (HR = 0.39; 95% CI: 0.20–0.75; $p = 0.004$), suggesting a lower likelihood of active disease.

Tumor stage at diagnosis was a powerful determinant of cancer status. Patients diagnosed at the distant stage had nearly double the hazard of having EOD compared to those with localized tumors (HR = 1.83; 95% CI: 1.30–2.56; $p < 0.001$). The regional stage also showed a significant association (HR = 1.41; 95% CI: 1.03–1.93; $p = 0.030$). Interestingly, patients with unknown stage had a markedly reduced hazard (HR = 0.08; 95% CI: 0.02–0.34; $p <$

0.001), though this likely reflects incomplete documentation or follow-up rather than a true clinical effect.

Table 3.9: Results of the Multivariate Cox Proportional Hazards Model for Cancer Status with Adjusted Hazard Ratios with 95% CI.

Variable	Hazard Ratio	95% CI	p-value
Age	1.01	1.002–1.015	0.007
Histology Category			
Adenocarcinoma NOS	1		
Mucinous Adenocarcinoma	0.95	0.659–1.355	0.760
Neuroendocrine & Carcinoid Tumors	0.56	0.323–0.957	0.034
Other Adenocarcinoma Subtypes	0.68	0.399–1.172	0.167
Others	0.39	0.204–0.747	0.004
Stage Category			
Localized and In-Situ	1		
Regional	1.41	1.034–1.928	0.030
Distant	1.83	1.302–2.560	< 0.001
Unknown	0.08	0.018–0.342	< 0.001
Regional Nodes			
Regional Nodes Negative	1		
No Nodes Examined	2.37	1.802–3.113	< 0.001
Regional Nodes Positive	1.61	1.263–2.041	< 0.001
Surgery			
No	1		
Yes	0.59	0.461–0.747	< 0.001
Radiation			

No	1		
Yes	0.69	0.550–0.876	0.002
Chemotherapy			
No	1		
Yes	0.29	0.234–0.364	< 0.001
Immunotherapy			
No	1		
Yes	0.66	0.430–1.013	0.057
Marital Status			
Married	1		
DIVORCED	1.50	0.906–2.489	0.115
SINGLE	1.77	1.312–2.378	< 0.001
UNKNOWN	1.73	1.226–2.455	0.002
WIDOWED	0.75	0.483–1.173	0.209

Lymph node involvement played a critical role. Patients with no nodes examined had more than double the hazard of EOD (HR = 2.37; 95% CI: 1.80–3.11; $p < 0.001$), and those with regional node positivity had a 61% higher hazard compared to those with negative nodes (HR = 1.61; 95% CI: 1.26–2.04; $p < 0.001$).

Regarding treatment, surgical resection significantly reduced the risk of EOD (HR = 0.59; 95% CI: 0.46–0.75; $p < 0.001$). Radiation therapy was also protective (HR = 0.69; 95% CI: 0.55–0.88; $p = 0.002$), and chemotherapy demonstrated a strong protective effect (HR = 0.29; 95% CI: 0.23–0.36; $p < 0.001$). While immunotherapy showed a trend toward decreased hazard (HR = 0.66; 95% CI: 0.43–1.01), it did not reach conventional statistical significance ($p = 0.057$).

Among social variables, being single (HR = 1.77; 95% CI: 1.31–2.38; $p < 0.001$) and having unknown marital status (HR = 1.73; 95% CI: 1.23–2.45; $p = 0.002$) were both associated with a significantly higher hazard of EOD compared to married patients. Being divorced showed a non-significant trend (HR = 1.50; $p = 0.115$), while widowed status was not statistically associated with the outcome ($p = 0.209$).

The Kaplan–Meier survival curves (Figure 3.4) compare the actual survival curve of training data with the predicted curve on the test set. The curves show a moderate alignment, indicating that the model's calibration was acceptable, though a slight overestimation of survival was observed in later time periods. This is likely due to a reduced number of EOD events at extended follow-up, leading to increased censoring.

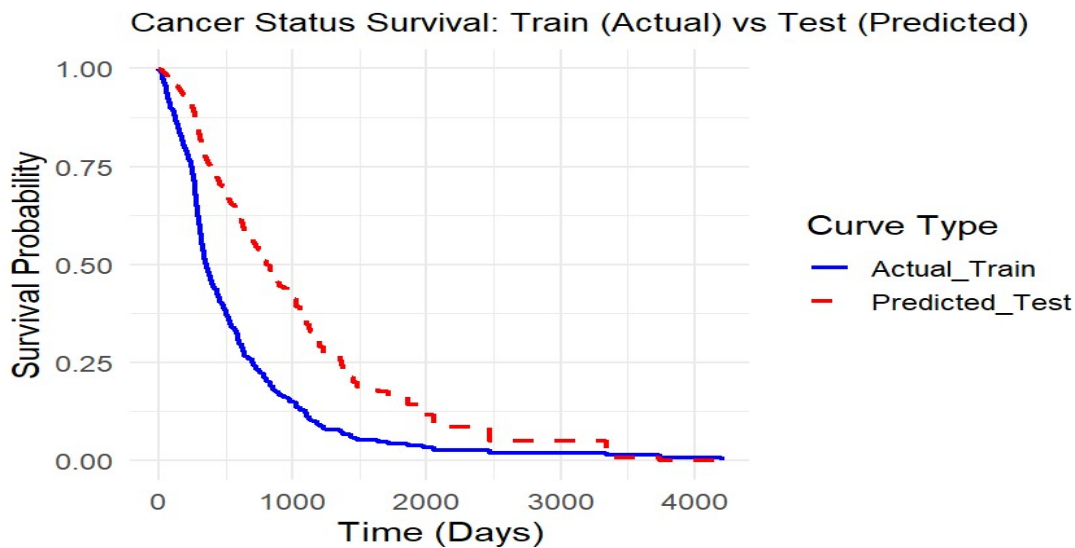


Figure 3.4: Observed vs. Predicted Kaplan–Meier Curves Cancer Status Survival curve

3.3.6 Multivariable Cox Proportional Hazards Analysis for Recurrence

This section presents the findings from the multivariable Cox regression analysis for recurrence-free survival. Initially, a full model including demographic, tumor, and treatment-related predictors was fitted. Following this, a backward stepwise selection procedure identified a final model with site category, stage at diagnosis, chemotherapy, and immunotherapy as significant predictors of recurrence (Table 3.10).

Tumor site in the rectum was significantly associated with increased recurrence risk. Compared to colon tumors, rectal tumors were associated with more than a fourfold higher hazard of recurrence (HR = 4.62; 95% CI: 1.74–12.26; $p = 0.002$). In contrast, tumors in the rectosigmoid junction showed no significant difference (HR = 1.07; $p = 0.873$).

The stage at diagnosis was another strong determinant of recurrence. Patients diagnosed at the distant stage had a sevenfold increased hazard of recurrence compared to those diagnosed at localized/in-situ stage (HR = 6.91; 95% CI: 1.23–38.62; $p = 0.028$). The regional stage was associated with a more than threefold increase in recurrence risk (HR = 3.09), but this did not reach statistical significance ($p = 0.095$).

Treatment factors were also informative. Receiving chemotherapy was significantly associated with an 83% reduction in recurrence risk (HR = 0.17; 95% CI: 0.03–0.99; $p =$

0.049), highlighting its therapeutic benefit in recurrence prevention. In contrast, immunotherapy was associated with a markedly increased risk of recurrence (HR = 7.11; 95% CI: 1.23–41.17; $p = 0.029$), potentially reflecting its use in more advanced or refractory cases.

Table 3.10: Results of the Multivariate Cox Proportional Hazards Model for Recurrence Status with Adjusted Hazard Ratios with 95% CI.

Predictors	Hazard Ratio (95% CI)	P-value
Stage at Diagnosis		
Localized and In-Situ	1.0 (reference)	
Distant	6.91 (1.23, 38.62)	0.028
Regional	3.09 (0.82, 11.59)	0.095
Unknown	0.09 (0.01, 0.85)	0.036
Primary Tumor Site		
Colon	1.0 (reference)	
Rectosigmoid	1.07 (0.48, 2.35)	0.873
Rectum	4.62 (1.74, 12.26)	0.002
chemotherapy		
No	1.0 (reference)	
Yes	0.17 (0.03, 0.99)	0.049
immunotherapy		
No	1.0 (reference)	
Yes	7.11 (1.23, 41.17)	0.029

Overall, the final model demonstrated moderate discrimination. The concordance index (C-index) was 0.72 (95% CI: 0.64–0.80) in the training set and 0.72 (95% CI: 0.45–0.98) in the test set, indicating reasonable predictive performance. The Kaplan–Meier survival curves (Figure 3.5) compare observed recurrence-free survival in the training data with predicted survival in the test data. The predicted curve tracked closely with the observed pattern throughout most of the follow-up period, reflecting satisfactory model calibration. The slight overestimation of survival at later time points is likely due to the low recurrence rate in the dataset, which results in a higher proportion of censored cases and reduced statistical precision in the tail of the curve. Despite this, the strong alignment between observed and predicted survival supports the model’s utility for recurrence risk prediction in CRC patients.

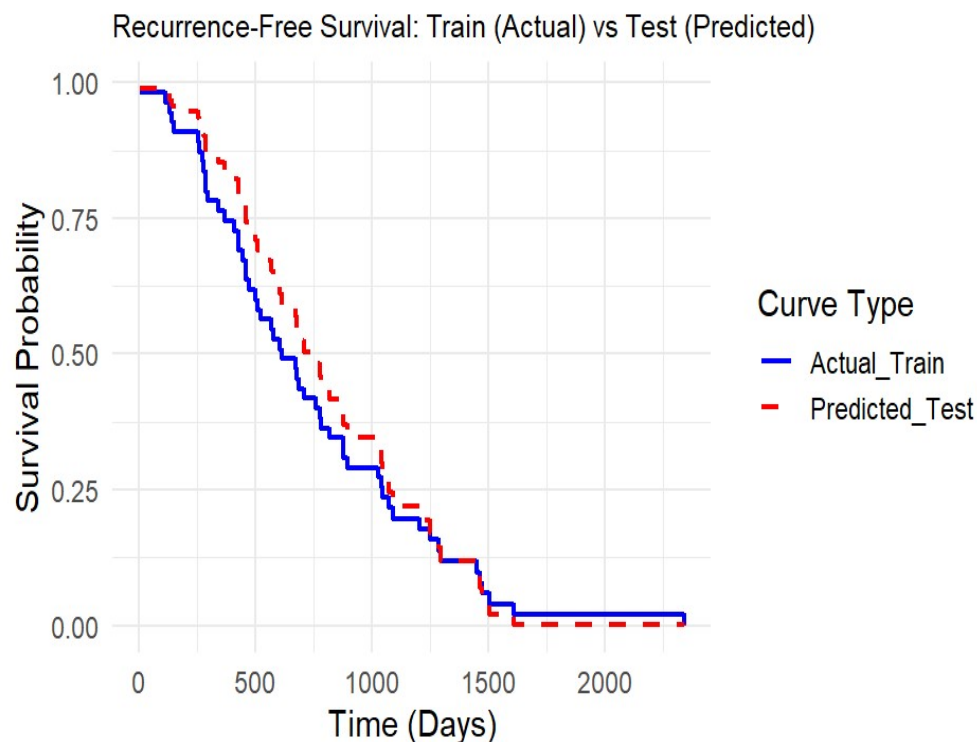


Figure 3.5: Observed vs. Predicted Kaplan–Meier Curves for Recurrence-Free Survival

3.3.7 Multivariable Cox Proportional Hazards Analysis for Overall Survival (Vital Status)

This section presents the results of the multivariable Cox proportional hazards model evaluating predictors of overall survival (Table 3.11). The model was fit using clinical and demographic covariates, including age, tumor characteristics, treatment modalities, and marital status. It demonstrated excellent discriminatory performance, achieving a C-index of 0.85 (95% CI: 0.83–0.87) in the training set and 0.88 (95% CI: 0.86–0.91) in the test set, indicating strong predictive capacity and generalizability.

Table 3.11: Results of the Multivariate Cox Proportional Hazards Model for Vital Status with Adjusted Hazard Ratios with 95% CI.

Variable	Hazard Ratio (95% CI)	P-value
Age	1.012 (1.002 , 1.022)	0.015
Marital status		
Married	1.0 (reference)	
Divorced	2.20 (1.11, 4.37)	0.024
Single	1.82 (1.12, 2.97)	0.016
Widowed	0.85 (0.43, 1.69)	0.650
Unknown	0.77 (0.34, 1.76)	0.536
Histology Behavior		
Mucinous Adenocarcinoma	2.44 (1.49, 3.98)	<0.001
Neuroendocrine & Carcinoid Tumors	0.59 (0.25, 1.39)	0.230
Other Adenocarcinoma Subtypes	1.73 (0.89, 3.36)	0.107
Others	0.70 (0.34, 1.44)	0.331
Regional Lymph Node Involvement		
Regional Nodes Negative	1.0 (reference)	
No Nodes Examined	2.68 (1.61, 4.46)	<0.001
Regional Nodes Positive	2.55 (1.61, 4.03)	<0.001
Stage at Diagnosis		
Localized and In-Situ	1.0 (reference)	
Distant	4.32 (2.31, 8.09)	<0.001
Regional	1.40 (0.76, 2.59)	0.284
Unknown	0.00 (0.00, 1,470,000)	0.989

Primary Tumor Site		
Colon	1.0 (reference)	
Rectosigmoid	0.62 (0.43, 0.90)	0.011
Rectum	0.70 (0.53, 0.91)	0.008
Appendix	0.65 (0.15, 2.81)	0.566
Surgery		
No	1.0 (reference)	
Yes	0.26 (0.18, 0.38)	<0.001
Chemotherapy		
No	1.0 (reference)	
Yes	0.41 (0.30, 0.56)	<0.001
Immunotherapy		
No	1.0 (reference)	
Yes	0.61 (0.35, 1.06)	0.081
Palliative		
No	1.0 (reference)	
Yes	1.28 (0.94, 1.73)	0.115

Age was significantly associated with mortality risk, with a hazard ratio (HR) of 1.012 (95% CI: 1.002–1.022; $p = 0.015$), indicating that each additional year of age increased the hazard of death by approximately 1.2%.

Among tumor characteristics, mucinous adenocarcinoma was associated with more than twice the hazard of death compared to adenocarcinoma NOS (HR = 2.44; 95% CI: 1.49–3.98; $p < 0.001$). Tumor site was also a significant predictor: patients with rectosigmoid (HR = 0.62; 95% CI: 0.43–0.90; $p = 0.011$) or rectum tumors (HR = 0.70; 95% CI: 0.53–0.91; $p = 0.008$) had significantly lower mortality risk compared to those with colon tumors.

Stage at diagnosis emerged as a powerful determinant of survival. Patients with distant-stage disease experienced a more than fourfold increase in the hazard of death (HR = 4.32; 95%

CI: 2.31–8.09; $p < 0.001$), whereas regional-stage disease was not statistically significant (HR = 1.40; $p = 0.284$). The "unknown" category had an implausibly low HR estimate due to data irregularities and was not interpretable ($p = 0.989$).

Lymph node involvement was significantly associated with worse outcomes. Patients with no nodes examined (HR = 2.68; 95% CI: 1.61–4.46; $p < 0.001$) and those with regional node positivity (HR = 2.55; 95% CI: 1.61–4.03; $p < 0.001$) were both at significantly higher risk of death compared to patients with negative nodes.

Regarding treatment modalities, surgical resection was strongly protective and associated with a 74% reduction in the hazard of death (HR = 0.26; 95% CI: 0.18–0.38; $p < 0.001$). Chemotherapy was also significantly protective (HR = 0.41; 95% CI: 0.30–0.56; $p < 0.001$). Although immunotherapy showed a trend toward reduced risk (HR = 0.61; 95% CI: 0.35–1.06), the association was not statistically significant ($p = 0.081$). Palliative care did not significantly predict overall survival (HR = 1.28; $p = 0.116$).

Marital status showed noteworthy associations: compared to married individuals, being divorced (HR = 2.20; 95% CI: 1.11–4.37; $p = 0.024$) or single (HR = 1.82; 95% CI: 1.12–2.97; $p = 0.016$) was significantly associated with higher mortality risk, suggesting potential psychosocial or support-related disparities.

The Kaplan-Meier survival curves (Figure 3.6) illustrate observed survival in the training set and predicted survival in the test set. The model-predicted survival closely followed the observed curve in the early periods but slightly overestimated survival in later time points. Nevertheless, the strong separation between the survival curves and consistent patterns underscore the model's good calibration and clinical utility in predicting overall survival in CRC patients.

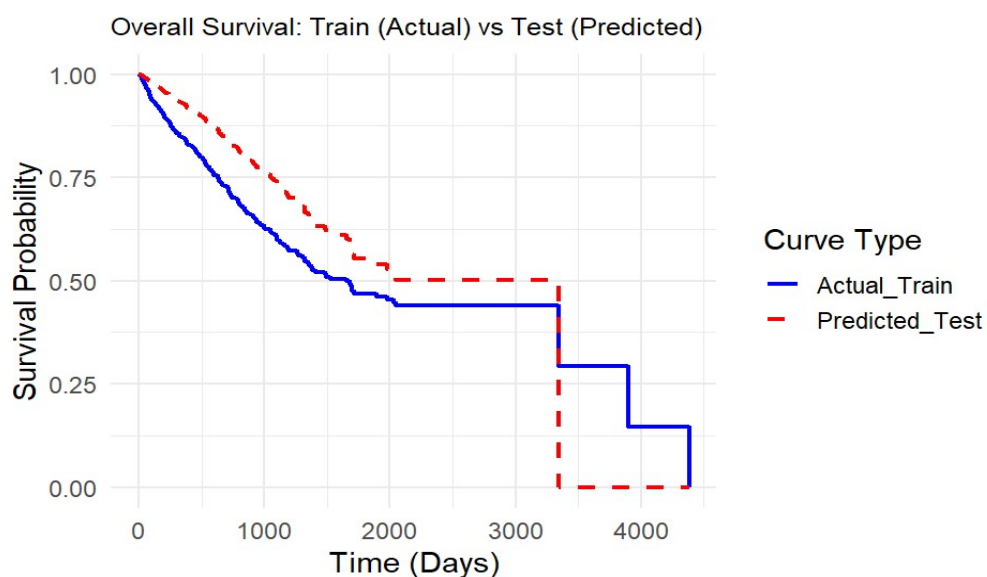


Figure 3.6: Observed vs. Predicted Kaplan–Meier Survival Curves for Overall Survival in CRC Patients.

3.4 Discussion

This study developed and validated multivariate LR and Cox proportional hazards models to predict CRC outcomes—specifically cancer status, recurrence, and overall survival—using data from 1,314 patients treated at King Fahd Specialist Hospital in Dammam, Saudi Arabia. The LR models showed strong predictive accuracy with AUCs of 0.80, 0.81, and 0.85 for recurrence, cancer status, and vital status, respectively. Similarly, Cox regression models demonstrated excellent discriminatory performance: for overall survival, the concordance index (C-index) was 0.85 (95% CI: 0.83–0.87) in the training set and 0.88 (95% CI: 0.86–0.91) in the test set; for recurrence-free survival, the C-index was 0.72 (95% CI: 0.64–0.80) in the training set and 0.72 (95% CI: 0.45–0.98) in the test set. For cancer status survival analysis, the C-index was 0.73 (95% CI: 0.71–0.75) in the training set and 0.68 (95% CI: 0.64–0.73) in the test set, reflecting acceptable discrimination and generalizability.

Across models, several clinical and demographic predictors were significantly associated with CRC outcomes. Age had a nuanced effect: while older patients had lower odds of recurrence (AOR = 0.97, $p = 0.003$), they exhibited higher risks of persistent disease (HR = 1.01, $p = 0.007$) and mortality (HR = 1.012, $p = 0.015$). These findings highlight the complex role of aging in cancer progression, treatment response, and survivorship.

Tumor characteristics were highly informative. Rectal tumors were associated with increased odds of cancer persistence (AOR = 2.09, $p = 0.002$) and a more than fourfold higher risk of recurrence (HR = 4.62, $p = 0.002$) compared to colon tumors. However, in terms of mortality, rectal and rectosigmoid tumors were linked to significantly lower hazard of death (HR = 0.70 and 0.62, respectively), suggesting different clinical behavior and possibly more aggressive treatment strategies for these tumor locations. Histological subtypes such as mucinous adenocarcinoma were linked to higher mortality (HR = 2.44, $p < 0.001$), while rare subtypes such as neuroendocrine and “others” were associated with a significantly lower hazard of persistent cancer status (HR = 0.56, $p = 0.034$ and HR = 0.39, $p = 0.004$, respectively).

Disease stage at diagnosis was among the most powerful predictors. Patients with distant-stage disease exhibited a sevenfold increased recurrence hazard (HR = 6.91, $p = 0.028$) and a fourfold higher risk of death (HR = 4.32, $p < 0.001$), highlighting the prognostic importance of early detection. Regional lymph node involvement was also a strong survival determinant: patients with no nodes examined or with regional node positivity had significantly higher mortality risk compared to those with negative nodes (HR = 2.68 and 2.55, respectively; both $p < 0.001$).

Lymph node evaluation was another critical predictor. Patients with no nodes examined had significantly worse outcomes in terms of cancer status (HR = 2.37, $p < 0.001$) and survival (HR = 2.68, $p < 0.001$). Regional node positivity similarly elevated risks of persistent disease

(HR = 1.61, $p < 0.001$) and mortality (HR = 2.55, $p < 0.001$), reaffirming the importance of nodal assessment in CRC management.

Treatment modalities had strong protective effects. Surgery significantly reduced the odds of disease presence (AOR = 0.15, $p < 0.001$) and hazards of recurrence (HR = 6.95, $p = 0.011$), cancer persistence (HR = 0.59, $p < 0.001$), and death (HR = 0.26, $p < 0.001$). Chemotherapy was a key protective factor across all models: it lowered the hazard of recurrence (HR = 0.17, $p = 0.049$), persistent disease (HR = 0.29, $p < 0.001$), and death (HR = 0.41, $p < 0.001$). Radiation therapy also showed protective effects for cancer persistence (HR = 0.69, $p = 0.002$). Immunotherapy, while associated with a higher hazard of recurrence (HR = 7.11, $p = 0.029$), showed a protective trend for cancer status (HR = 0.66, $p = 0.057$) and survival (HR = 0.61, $p = 0.081$), possibly reflecting its use in advanced-stage or immunoresponsive patients.

Marital status was a significant social determinant: single and divorced patients had higher mortality risk compared to married counterparts (HR = 1.82 and 2.20, respectively), reinforcing the protective effects of social support documented in prior literature. This association underscores the importance of psychosocial factors in CRC prognosis and supports the integration of social determinants into predictive modeling.

Overall, the models demonstrated robust predictive performance and strong calibration, as evidenced by ROC curves and Kaplan-Meier survival comparisons. Particularly, the Cox model for overall survival achieved high concordance, and the predicted survival curves closely aligned with observed data.

Despite these strengths, the study has limitations. The recurrence rate was relatively low (5.5%), which, although reflective of real-world outcomes, may limit statistical power and precision, especially in the Cox model for recurrence. Additionally, certain rare categories (e.g., appendix tumors, neuroendocrine subtypes) were underrepresented, limiting the ability to draw firm conclusions about their prognostic implications. Moreover, the data originated from a single center in Saudi Arabia, which may constrain the generalizability of the findings to broader populations. Potential unmeasured confounders—such as comorbidities, genomic data, or socioeconomic factors—were not included, which may affect model completeness.

Nevertheless, this study provides one of the first comprehensive statistical modeling frameworks for CRC outcomes in a Middle Eastern clinical setting. The models offer valuable clinical insights and hold potential for incorporation into digital tools to assist in prognosis, treatment planning, and patient communication.

In conclusion, the findings affirm the utility of LR and Cox proportional hazards models for predicting key CRC outcomes, including recurrence, cancer status, and survival. The identified predictors—age, tumor site, disease stage, lymph node involvement, treatment modalities, and marital status—align with and extend existing literature. These models form

a foundation for future predictive analytics in CRC and support the development of risk stratification tools tailored to regional populations. Further research should explore integration with ML techniques, external validation in other cohorts, and expansion to include genomic and lifestyle data for more precise individualized care.

4. Machine Learning Models for Predicting Key Outcomes in CRC Patients in Saudi Arabia

4.1 Background

This chapter presents a comparative evaluation of multiple ML algorithms for predicting three critical clinical outcomes in CRC patients: recurrence, cancer status at follow-up, and overall survival (vital status). Using retrospective data from 1,314 patients treated at King Fahd Specialist Hospital in Dammam, Saudi Arabia, five widely adopted ML techniques were implemented and assessed. These included Random Forest (RF), Support Vector Machines with both linear and radial kernels (svmLinear and svmRadial), Extreme Gradient Boosting (XGBLinear), and Artificial Neural Networks (ANN). These algorithms were selected for their demonstrated ability to model complex relationships in clinical data and their effectiveness in medical prediction tasks.

The primary aim of this chapter is to identify the most accurate and generalizable models for each of the targeted CRC outcomes. By doing so, the study seeks to contribute to the development of reliable, data-driven clinical decision-support tools that can assist oncologists in patient risk stratification, treatment planning, and follow-up management. Accurate prediction of recurrence and survival outcomes is essential in delivering timely interventions and improving personalized cancer care.

The chapter begins by detailing the methodological approach used to develop, train, and evaluate the predictive models, emphasizing the steps taken to ensure reproducibility and fairness in comparison. It then briefly revisits the dataset introduced earlier in the thesis, grounding the analysis in the real-world clinical context of CRC patients in Saudi Arabia. Following this, attention is given to the process of selecting the most informative features for each model, highlighting the role of clinical relevance and data-driven refinement. The comparative analysis unfolds by examining each of the three outcomes—cancer status, recurrence, and vital status—in turn, presenting model performance results alongside visualizations such as ROC curves and discussing the features that consistently emerged as important across algorithms. The chapter culminates in a reflection on the key findings, their clinical implications, and the broader potential of ML to support personalized cancer care.

4.2 Methods

To predict CRC outcomes using ML, this study adopted a systematic data science pipeline that ensured methodological rigor and reproducibility. All analyses were conducted using R (version 4.4.2), a powerful open-source statistical programming environment widely utilized

for clinical research and predictive modeling. The workflow encompassed data preprocessing, model development, hyperparameter tuning, and performance evaluation.

In the preprocessing stage, the dataset was cleaned to address inconsistencies, and categorical variables were appropriately encoded to ensure compatibility with ML algorithms. Rather than using a fixed train-test split, the analysis employed a nested cross-validation framework to ensure rigorous model evaluation and reduce overfitting.

A range of ML techniques was applied to predict each of the three CRC outcomes: recurrence, cancer status, and vital status. The selected models included RF, SVM) with both linear and radial kernels, Extreme Gradient Boosting using the linear booster (XGBLinear), and Artificial Neural Networks (ANN). These models were chosen for their proven effectiveness in clinical prediction tasks and their ability to model complex, nonlinear relationships in health data.

Model training and tuning were conducted using nested cross-validation within the full dataset to optimize generalizability and robustness. Hyperparameters for each model were selected based on cross-validated performance metrics. Final model performance was assessed using a bootstrap-based evaluation strategy on out-of-bag samples over 500 simulations. Key performance metrics included accuracy, sensitivity, specificity, PPV, NPV, and the AUC.

This comprehensive and structured approach allowed for consistent comparisons across models and outcomes, ultimately identifying the most effective ML techniques for CRC outcome prediction.

4.2.1 Dataset

For a comprehensive description of the dataset utilized in this study - including information on the study population, outcome variables (recurrence, cancer status, and vital status), and predictor variables - refer to Sections 3.2.1 (Study Design and Participants), 3.2.2 (Outcome Variables), and 3.2.3 (Predictor Variables) in the previous chapter.

4.2.2 Feature Selection for Prognostic Modeling

In the context of predicting CRC outcomes, careful selection of prognostic features is fundamental to developing clinically relevant and robust predictive models. Rather than relying on arbitrary variable inclusion, this study grounded its feature set in prior literature and clinical insight. The chosen predictors represent key domains of CRC prognosis - encompassing patient demographics, tumor characteristics, and treatment-related interventions - all of which are widely recognized as influential in determining patient outcomes.

Given the structured nature of the dataset ($n = 1,314$) and the need to balance predictive power with interpretability, we adopted an embedded feature selection strategy within the ML workflow. Specifically, backward elimination was employed within each fold of the nested cross-validation framework. This iterative method starts with the complete set of predictors and sequentially removes the least informative variables, using the AUC as the optimization criterion. The process continues until no further gain in performance is observed, yielding a parsimonious model tailored to each learning algorithm and outcome.

This approach offers two distinct advantages. First, by integrating feature selection into the training process, it mitigates the risk of overfitting, particularly important in clinical datasets where redundancy and multicollinearity are common. Second, it enhances model transparency, highlighting which variables most consistently contribute to predictive performance.

The final models varied slightly by algorithm and outcome but consistently retained clinically meaningful variables such as tumor site, stage at diagnosis, regional lymph node involvement, surgery, chemotherapy, immunotherapy, and patient age. These features are well-aligned with established CRC prognostic indicators, lending further support to the validity of the data-driven selection process.

The next section outlines the evaluation framework and performance metrics used to validate the resulting models and ensure their applicability in real-world clinical settings.

4.2.3 Validation Strategy and Performance Evaluation Metrics

In this study, robust validation strategies were employed to evaluate and compare the performance of multiple ML models in predicting CRC outcomes—specifically cancer status, recurrence and vital status. The analysis was conducted using a nested cross-validation framework combined with feature selection and bias-corrected performance estimation through bootstrapping.

Nested Cross-Validation and Feature Selection

To prevent overfitting and promote generalizability, a 10-fold nested cross-validation scheme was adopted. For each outer fold, the data were divided into training and testing subsets, with the test portion used exclusively for model evaluation. Within each outer training set, an inner 5-fold cross-validation loop was employed to perform hyperparameter tuning using a grid search strategy across predefined sets of values.

After hyperparameter tuning, feature selection was conducted on the outer training data using a backward elimination approach. In each iteration, the least impactful variable was removed, and the model was retrained until an optimal subset of predictors was identified, maximizing the Area Under the ROC Curve (AUC). This process ensured that the selected features provided the highest predictive power while mitigating the risk of overfitting—a crucial

consideration in biomedical datasets characterized by high dimensionality and potential noise.

The outer loop evaluated the generalization performance of the models using independent test folds. The model achieving the highest sensitivity across all outer folds was preserved for subsequent bootstrap validation and further analysis.

Handling Class Imbalance

Class imbalance was a critical issue, particularly for the recurrence outcome, which had a low prevalence (~5.5%) in the dataset. To address this, two different resampling strategies were implemented at different stages of model development. SMOTE was employed during the nested validation and hyperparameter optimization process to increase the representation of the minority class in training data, allowing models to develop more stable and accurate decision boundaries. In contrast, up sampling was applied in the final 500 bootstrap validation simulations, where minority class instances were randomly duplicated to match the majority class. This ensured balanced training while preserving the integrity of test evaluations without introducing synthetic data. Together, these techniques mitigated class imbalance and enhanced the models' ability to detect rare outcomes reliably.

Model Selection and Implementation

The selected ML models comprised RF, svmLinear, svmRadial, xgbLinear and ANN. Each algorithm was integrated into the same nested validation and feature selection pipeline to ensure a fair comparison.

The final models were stored with their corresponding hyperparameters and important variables for reproducibility and further analysis. Model performance was reported based on both in-sample and out-of-sample (bootstrap test) metrics, allowing for a rigorous and transparent comparison.

4.3 Result

4.3.1 Cancer Status Prediction

Five ML algorithms were applied to estimate cancer status at the final follow-up of CRC patients, using a 10-fold nested cross-validation approach with 500 bootstrap simulations. Cancer status was defined as either No Evidence of Disease (NED) or Evidence of Disease (EOD). The modeling pipeline involved data preprocessing, class balancing using SMOTE and up-sampling, and embedded feature selection to enhance model performance and interpretability.

Of all the evaluated models, XGBLinear showed the best training performance (Table 4.1), recording an accuracy of 0.7504 and an AUC of 0.8388. The ANN model performed comparably, reaching 0.7397 in accuracy and 0.8342 in AUC, reflecting high discriminative capability. The RF model stood out for its high specificity (0.9353) and PPV (0.9438), though it exhibited a lower sensitivity of 0.5929, suggesting a more conservative approach in identifying patients with evidence of disease. The svmRadial achieved the highest sensitivity (0.7212) among all models, making it effective in capturing positive cases, albeit with a lower specificity (0.7328). All models recorded AUC values above 0.80, underscoring their effectiveness in distinguishing between patients with and without disease evidence.

Table 4.1: Training Set Evaluation Metrics for Machine Learning Models Predicting Cancer Status

Model	Accuracy	Specificity	Sensitivity	NPV	PPV	AUC
RF	0.7139	0.9353	0.5929	0.5564	0.9438	0.8247
ANN	0.7397	0.8858	0.6600	0.5871	0.9137	0.8342
svmLinear	0.6971	0.9159	0.5776	0.5421	0.9264	0.8034
svmRadial	0.7253	0.7328	0.7212	0.5893	0.8318	0.8036
xgbLinear	0.7504	0.9052	0.6659	0.5966	0.9279	0.8388

Validation on the bootstrap out-of-bag samples further highlighted the robustness of the models. (Table 4.2). with all achieving AUC values above 0.78. The RF model recorded the highest specificity (0.9204) and PPV (0.9290), reflecting a conservative approach that favors minimizing false positives, though at the cost of lower sensitivity (0.5689). The ANN and XGBLinear models demonstrated the best balance between accuracy (0.7200 and 0.7202, respectively), sensitivity (0.6492 and 0.6495), and specificity (0.8492 and 0.8494), while maintaining high AUC values (0.7944 and 0.7854). The svmLinear model showed strong specificity (0.8935) and PPV (0.9096), paired with moderate sensitivity (0.5886), supporting its use in scenarios requiring high certainty in positive predictions. The svmRadial model again achieved the highest sensitivity (0.6934), consistent with its training behavior, though its specificity was the lowest (0.7352), indicating a trade-off in favor of recall. These results emphasize the ANN and XGBLinear models' strong generalization performance, while highlighting the distinct strengths of RF and SVM-based models in different clinical decision-making contexts.

Table 4.2: Validation Set Evaluation Metrics for Machine Learning Models Predicting Cancer Status

Model	Accuracy	Specificity	Sensitivity	NPV	PPV	AUC
RF	0.6931	0.9204	0.5689	0.5385	0.9290	0.8008
ANN	0.7200	0.8492	0.6492	0.5703	0.8870	0.7944
svmLinear	0.6967	0.8935	0.5886	0.5440	0.9096	0.7960
svmRadial	0.7082	0.7352	0.6934	0.5659	0.8282	0.7847
xgbLinear	0.7202	0.8494	0.6495	0.5694	0.8880	0.7854

When comparing overall performance across training and validation phases, the ANN and XGBLinear models emerged as the most balanced and reliable classifiers for predicting cancer status. Both consistently demonstrated strong accuracy, AUC, and sensitivity metrics, indicating effective generalization to unseen data. ANN showed a slightly better validation AUC (0.7944) compared to XGBLinear (0.7854), while maintaining comparable accuracy and predictive values, making it particularly effective in identifying patients with evidence of disease.

RF consistently offered the highest specificity and PPV across both datasets, favoring precision and minimizing false positives. However, its relatively low sensitivity suggests a tendency to under-identify true positive cases, which may limit its utility in screening settings where recall is critical. Conversely, the svmRadial prioritized sensitivity, achieving the highest recall in both training (0.7212) and validation (0.6934), though at the cost of lower specificity, rendering it appropriate for scenarios where elevated sensitivity is essential.

The svmLinear maintained strong specificity and PPV across both sets, positioning it as a conservative model that performs well in confirming positive cases, albeit with modest sensitivity and AUC. Overall, ANN and XGBLinear demonstrated the most robust and generalizable performance, while RF and SVM variants showed potential for use in context-specific clinical applications depending on the prioritization of precision versus recall.

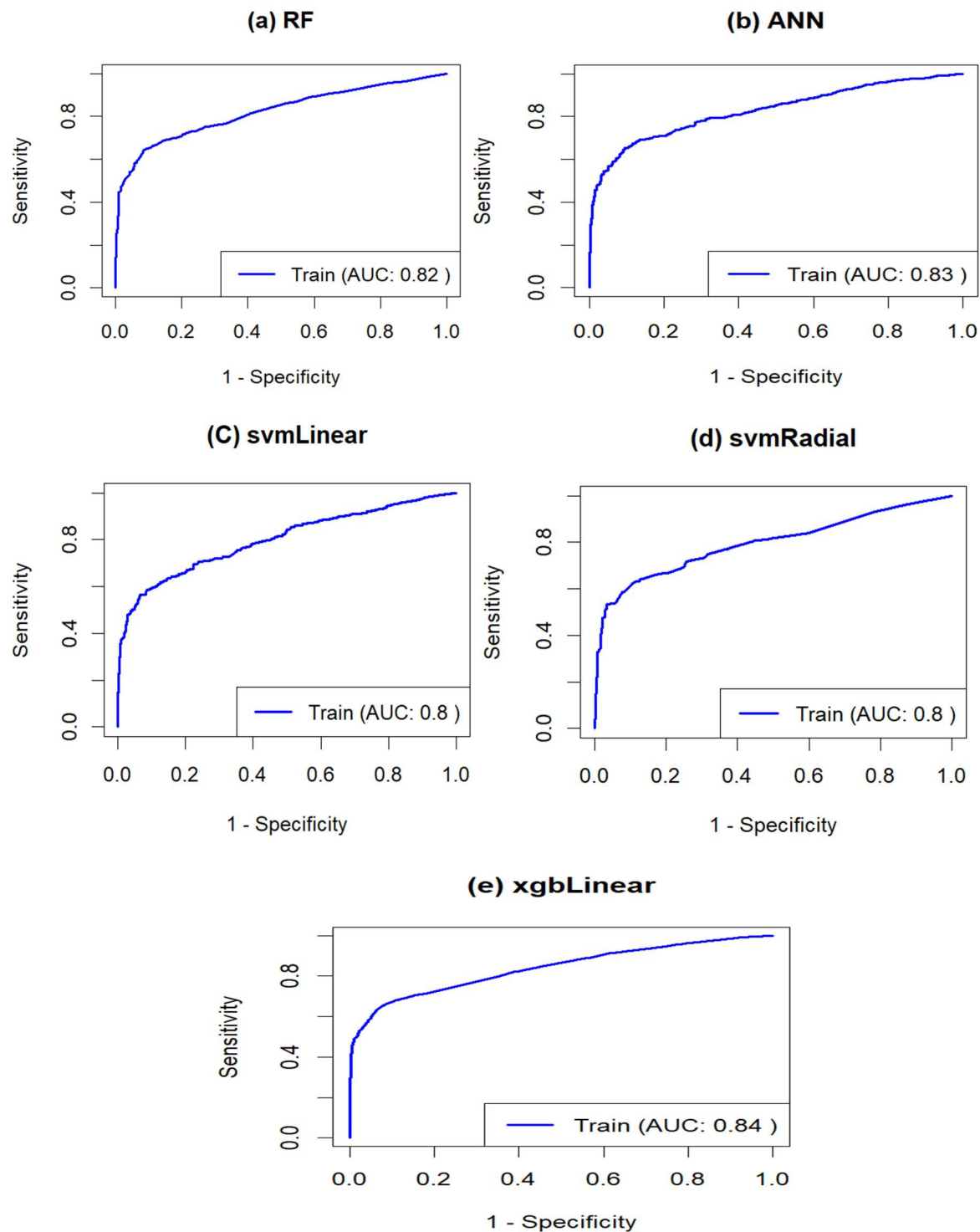


Figure 4.1: ROC Curves of Machine Learning Models for Predicting Cancer Status in CRC Patients (Training Set)

Figure 4.1 presents the ROC curves for each ML model used to predict cancer status in CRC patients, evaluated on the training set. These curves reflect how each model balances sensitivity and specificity at different threshold levels, with the AUC indicating the model's ability to distinguish between outcome classes.

XGBLinear recorded the top AUC value of 0.84, reflecting its superior capacity to distinguish between patients with and without evidence of disease. The ANN model closely followed with an AUC of 0.83, maintaining a smooth and well-distributed curve indicative of robust performance. The RF model also performed well (AUC = 0.82), especially notable for its strong specificity and positive predictive value.

Both svmLinear and svmRadial models demonstrated similar AUC values of 0.80. However, their ROC curves suggest different operating characteristics—svmRadial emphasized higher sensitivity at the expense of specificity, while svmLinear maintained stronger specificity and a more conservative classification approach.

Overall, the ROC curves confirm that all five models possess strong discriminative power (AUC > 0.80), with XGBLinear and ANN demonstrating the best balance between recall and precision. These findings align with the numerical results in Tables 4.1 and 4.2, supporting their potential use in clinical settings for early identification of patients with disease evidence.

Feature selection for cancer status prediction revealed a consistent set of prognostic variables retained across multiple ML models (Table 4.3). Variables such as stage category, site category, regional lymph node involvement, and radiation were selected by all five algorithms, underscoring their central role in predicting cancer status. Treatment-related features, including surgery, palliative care, and immunotherapy, were also frequently retained, particularly by ensemble and neural network-based models. In contrast, demographic features like age, gender, and marital status appeared less frequently across models, indicating that their predictive value might vary depending on the algorithm or its interaction with other predictors. Overall, the selected features align well with known clinical indicators of disease status, supporting the validity of the data-driven selection process employed within each modeling framework.

Table 4.3: Features Retained in Final Machine Learning Models for Cancer Status Prediction

Variable	RF	ANN	svmLinear	svmRadial	XgbLinear
Gender	✓	✓	✓		
Age			✓		
Marital		✓			
Surgery	✓	✓	✓	✓	

Radiation	✓	✓	✓	✓	✓
Chemotherapy	✓	✓		✓	
Immunotherapy	✓	✓	✓		
Palliative	✓	✓	✓	✓	
Regional Lymph Node	✓	✓	✓	✓	✓
Stage at Diagnosis	✓	✓	✓	✓	✓
Primary Tumor Site	✓	✓	✓	✓	✓

A check mark (✓) indicates that the corresponding variable was selected in the final version of each model. Blank cells denote features that were excluded during the embedded feature selection process.

4.3.2 Recurrence Prediction

To predict disease recurrence among CRC patients, Five ML algorithms were developed and assessed through a 10-fold nested cross-validation scheme with a 500-iteration bootstrap validation. Recurrence was defined as the reappearance of cancer within the follow-up period. Given the low prevalence of recurrence in the dataset, class imbalance was addressed using SMOTE during model development and up-sampling in the final evaluation phase.

Among all models, the XGBLinear algorithm achieved the highest AUC on the training set (Table 4.4), reaching 0.8172, along with a sensitivity of 0.8333 and an accuracy of 0.6804. The Artificial Neural Network (ANN) followed closely, recording the highest sensitivity (0.8611) and an AUC of 0.7994. However, its performance was accompanied by reduced specificity (0.6264) and accuracy (0.6393), suggesting a tendency toward over-predicting recurrence. The svmLinear model exhibited a similar pattern, achieving an AUC of 0.7891 with a sensitivity of 0.8194, but slightly lower specificity (0.6216) and accuracy (0.6324). RF demonstrated a balanced profile, achieving 0.7062 in accuracy, 0.8056 in sensitivity, and an AUC score of 0.7663, positioning it as a strong candidate for general-purpose classification. The svmRadial model delivered the most balanced trade-off between specificity (0.7174) and sensitivity (0.6806), although with a comparatively lower AUC of 0.7764. Despite all models showing modest PPV—due to the highly imbalanced outcome—all achieved high NPV (>0.97), which is essential for confidently ruling out recurrence in low-risk patients.

Table 4.4: Training Set Evaluation Metrics for Machine Learning Models Predicting CRC Recurrence

Model	Accuracy	Specificity	Sensitivity	NPV	PPV	AUC
RF	0.7062	0.7005	0.8056	0.9842	0.1349	0.7663
ANN	0.6393	0.6264	0.8611	0.9873	0.1179	0.7994
svmLinear	0.6324	0.6216	0.8194	0.9834	0.1115	0.7891
svmRadial	0.7154	0.7174	0.6806	0.9748	0.1225	0.7764
xgbLinear	0.6804	0.6715	0.8333	0.9852	0.1282	0.8172

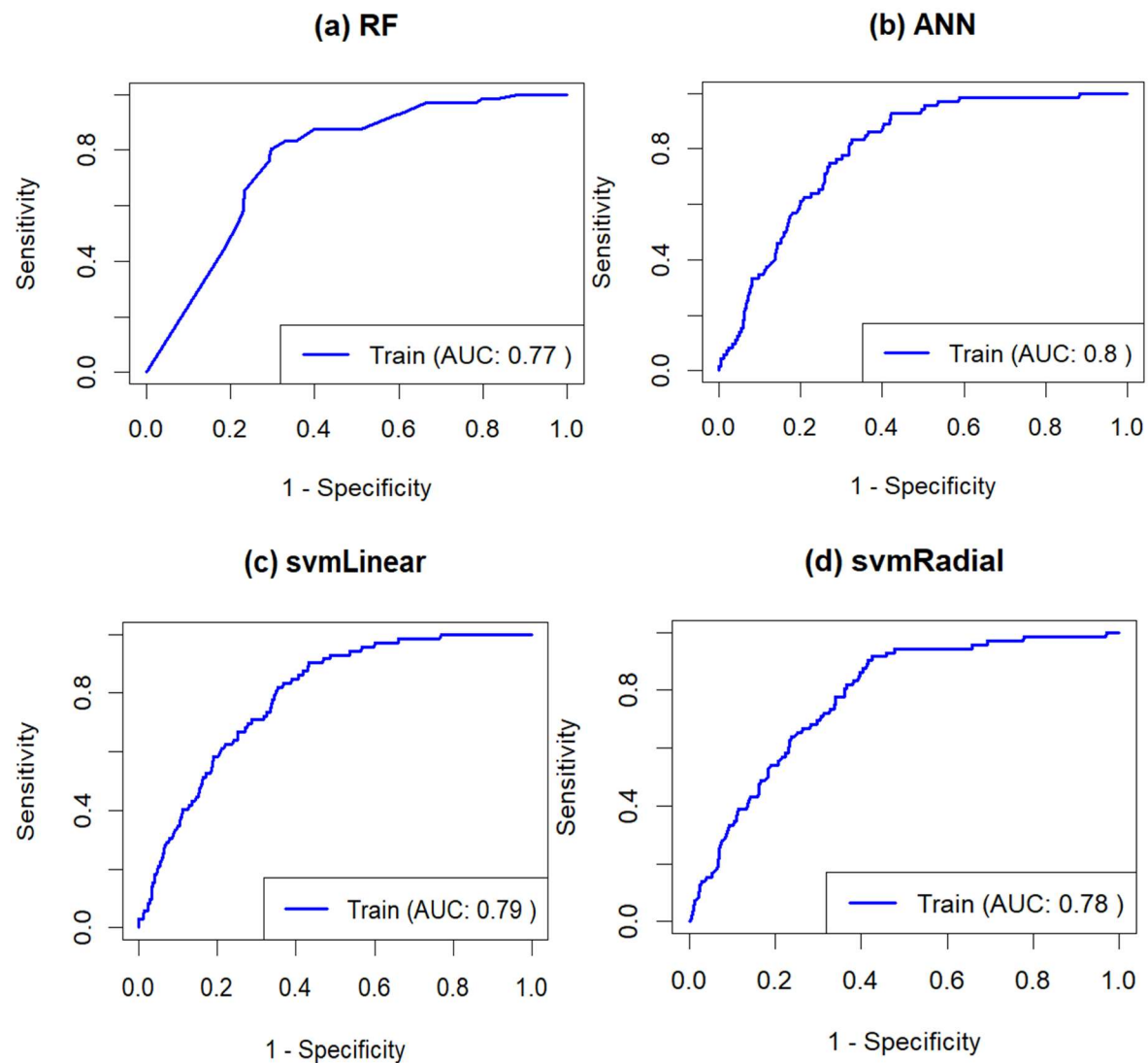
Validation on the bootstrap out-of-bag samples (Table 4.5) confirmed the overall trends seen during training. The XGBLinear model maintained its superior performance, achieving the highest AUC (0.7641) and demonstrating strong sensitivity (0.7693) with good specificity (0.6527), underscoring its ability to generalize to unseen data. The RF model followed closely, with an AUC of 0.7396 and the highest sensitivity (0.7717), suggesting high utility in identifying patients at risk of recurrence. ANN and svmLinear models both maintained strong sensitivity (0.7352 and 0.7562, respectively) but exhibited lower specificity and the lowest PPVs, reflecting their bias toward detecting positives. Meanwhile, svmRadial achieved the highest specificity (0.7277) among all models in the validation phase, though it had the lowest sensitivity (0.6429), favoring a more conservative strategy.

Table 4.5: Validation Set Evaluation Metrics for Machine Learning Models Predicting CRC Recurrence

Model	Accuracy	Specificity	Sensitivity	NPV	PPV	AUC
RF	0.6496	0.6425	0.7717	0.9798	0.1114	0.7396
ANN	0.6199	0.6134	0.7352	0.9759	0.0982	0.7361
svmLinear	0.6052	0.5964	0.7562	0.9768	0.0983	0.7302
svmRadial	0.7231	0.7277	0.6429	0.9728	0.1184	0.7348
xgbLinear	0.6590	0.6527	0.7693	0.9798	0.1135	0.7641

When comparing overall performance across both training and validation phases, the XGBLinear model emerged as the most reliable and generalizable classifier for recurrence prediction. It consistently achieved the highest AUC values (0.8172 and 0.7641) while maintaining a strong balance between sensitivity and specificity. The RF model also performed robustly, particularly in maximizing recall (sensitivity = 0.7717), making it highly effective in capturing recurrence cases. The ANN model was notable for its high sensitivity

but demonstrated lower specificity and PPV, which may lead to more false positives. Both SVM models maintained moderate AUCs—svmLinear peaking at 0.7891 during training and svmRadial achieving 0.7348 on validation—reflecting acceptable but less consistent performance compared to ensemble and neural network models. They also exhibited distinct trade-offs: svmLinear emphasized recall, while svmRadial prioritized specificity. Collectively, these findings highlight the strengths of XGBLinear and RF as leading models for recurrence prediction, with ANN offering a high-sensitivity alternative suitable for risk screening scenarios.



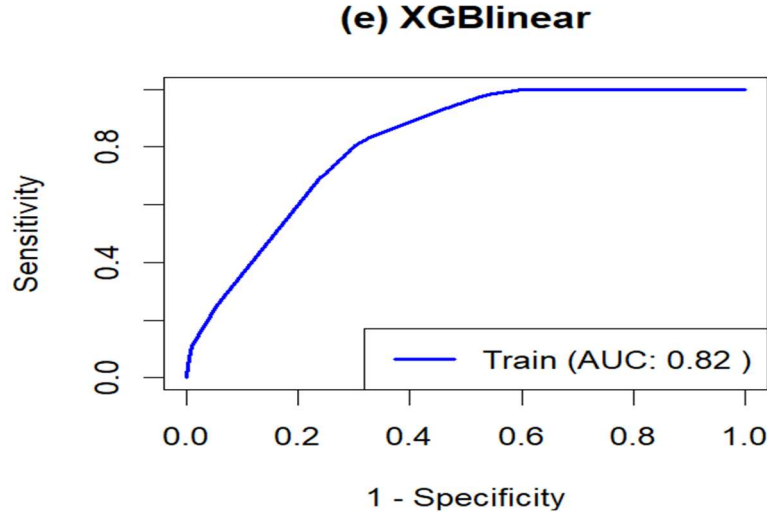


Figure 4.2: ROC Curves of Machine Learning Models for Predicting CRC Recurrence in the Training Set

Figure 4.2 presents the ROC curves for each ML model used to predict CRC recurrence, evaluated on the training set. The XGBLinear model achieved the highest AUC (0.82), demonstrating its strong capability to differentiate patients based on recurrence status. The ANN model followed with an AUC of 0.80 and displayed a smooth, steadily rising curve reflective of its high sensitivity. RF achieved a slightly lower AUC (0.77) but maintained a sharp ascent in the low false-positive region, aligning with its strong recall performance. The SVM-based models also performed reasonably well, with svmLinear and svmRadial achieving AUCs of 0.79 and 0.78, respectively. Notably, svmRadial exhibited a steeper rise in sensitivity at early thresholds, suggesting stronger early recall, whereas svmLinear showed a more gradual slope consistent with conservative classification behavior. These patterns support the quantitative results in Table 4.4, confirming that XGBLinear and ANN models offer the most balanced discriminative performance, while RF and SVMs may be preferable in settings prioritizing sensitivity or specificity.

Feature selection results for recurrence prediction are summarized in Table 4.6. Across the five models, surgery and chemotherapy were the only predictors retained consistently by all algorithms, highlighting their critical role in determining the likelihood of disease recurrence. Stage category and site category were selected by RF, ANN, and XGBLinear, reaffirming their importance in risk stratification. Interestingly, models such as svmLinear and svmRadial included a broader mix of demographic and clinical variables, such as age, marital status, and histology category, reflecting their sensitivity to feature interactions.

Table 4.6: Features Retained in Final Machine Learning Models for CRC Recurrence Prediction

Variable	RF	ANN	svmLinear	svmRadial	xgbLinear
Gender			✓	✓	
Age		✓	✓	✓	
Marital status			✓	✓	✓
Surgery	✓	✓	✓	✓	✓
Radiation			✓		
Chemotherapy	✓	✓	✓	✓	✓
Stage at Diagnosis	✓	✓			✓
Primary Tumor Site	✓	✓			✓
Histology – Behavior			✓		✓

A check mark (✓) indicates that the corresponding variable was selected in the final version of each model. Blank cells denote features that were excluded during the embedded feature selection process.

4.3.3 Vital Status Prediction

To predict vital status among CRC patients (alive vs. deceased), Five ML algorithms were constructed and evaluated through a 10-fold nested cross-validation design with 500 bootstrap iterations for external validation. Vital status was defined based on the last available clinical follow-up. The modeling process followed the same pipeline used in previous outcomes, including preprocessing, resampling to mitigate class imbalance, and embedded feature selection.

Within the training dataset, XGBLinear outperformed other models across most evaluation metrics (Table 4.7), attaining an AUC score of 0.9563 and an accuracy rate of 0.8881, and a balanced sensitivity (0.8811) and specificity (0.8910). The RF and svmRadial models followed closely with identical accuracy (0.8394) and similar AUCs (0.8917 and 0.8778, respectively), with strong NPVs (>0.90) and moderate PPVs. The ANN algorithm also showed strong results, with an AUC of 0.8499, while preserving a good balance between sensitivity (0.7649) and specificity (0.8382). Interestingly, svmLinear showed the highest

sensitivity (0.8243) aside from XGBLinear, but with lower specificity (0.7422) and AUC (0.8365), suggesting a model more sensitive to detecting deceased cases at the cost of some false positives. All models displayed strong NPVs, indicating reliability in identifying patients likely to remain alive.

Table 4.7: Training Set Evaluation Metrics for Machine Learning Models Predicting Vital Status in CRC Patients

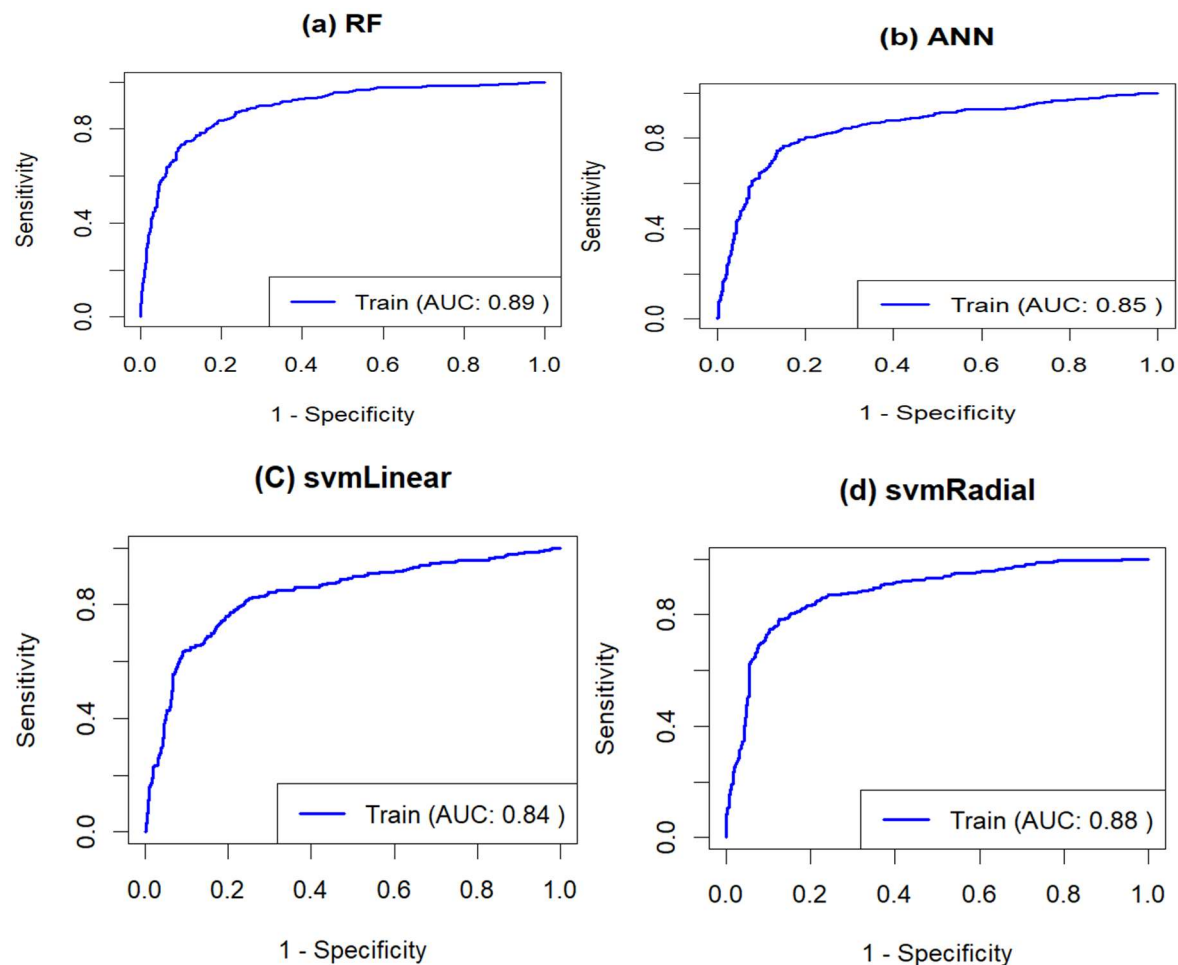
Model	Accuracy	Specificity	Sensitivity	NPV	PPV	AUC
RF	0.8394	0.8641	0.7804	0.9041	0.7056	0.8917
ANN	0.8166	0.8382	0.7649	0.8952	0.6637	0.8499
svmLinear	0.7664	0.7422	0.8243	0.9101	0.5717	0.8365
svmRadial	0.8394	0.8630	0.7829	0.9050	0.7047	0.8778
xgbLinear	0.8881	0.8910	0.8811	0.9472	0.7715	0.9563

Validation results emphasized the comparative advantages and limitations of each algorithm when applied to unseen data (Table 4.8). The RF model yielded the top AUC value of (0.8439), along with strong specificity (0.8431) and sensitivity (0.7276), underscoring its balanced and robust classification performance. The ANN model also performed competitively, with a high sensitivity (0.7515) and AUC (0.8357), making it effective in identifying deceased cases. The svmLinear model prioritized sensitivity (0.7308) and demonstrated solid discrimination (AUC = 0.8271), while the svmRadial model maintained similar performance with a slightly lower AUC (0.8211). In contrast, the XGBLinear model, despite its strong training performance, exhibited the lowest validation AUC (0.7971) and sensitivity (0.6589), indicating possible overfitting. Across all models, NPV values remained high (>0.85), highlighting their reliability in correctly identifying patients who are likely to remain alive.

Table 4.8: Validation Set Evaluation Metrics for Machine Learning Models Predicting Vital Status in CRC Patients

Model	Accuracy	Specificity	Sensitivity	NPV	PPV	AUC
RF	0.8090	0.8431	0.7276	0.8809	0.6600	0.8439
ANN	0.7957	0.8142	0.7515	0.8866	0.6289	0.8357
svmLinear	0.7759	0.7947	0.7308	0.8762	0.5975	0.8271
svmRadial	0.7938	0.8236	0.7218	0.8769	0.6297	0.8211
xgbLinear	0.7674	0.8128	0.6589	0.8506	0.5957	0.7971

When evaluating performance across both training and validation sets, the XGBLinear model emerged as the top performer in the training phase, with an outstanding AUC of 0.9563, along with high sensitivity (0.8811) and strong specificity (0.8910). However, it showed a notable performance drop during validation (AUC = 0.7971), suggesting potential overfitting. In contrast, the RF model demonstrated strong and stable performance across both phases, with a validation AUC of 0.8439 and well-balanced sensitivity (0.7276) and specificity (0.8431), highlighting its reliability and generalizability. The ANN model also maintained consistent results, achieving a validation AUC of 0.8357 and balanced sensitivity (0.7515) and specificity (0.8142), reinforcing its applicability in detecting high-risk patients. The svmLinear and svmRadial models offered modest AUCs (0.8271 and 0.8211, respectively) and preserved good NPVs across both sets, though their lower PPVs and sensitivity-specificity trade-offs indicate more limited use cases. Notably, svmLinear emphasized recall, while svmRadial prioritized a more balanced classification. Overall, RF and ANN demonstrated the most robust and generalizable performance for vital status prediction, with XGBLinear showing high training capacity but reduced external performance, indicating the importance of validation in real-world applications.



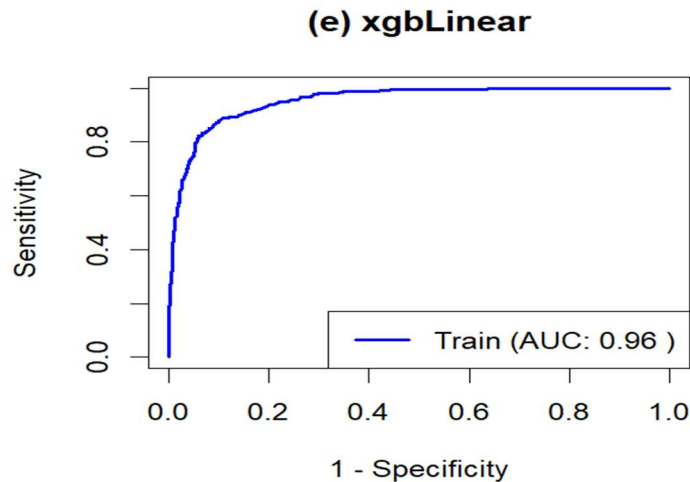


Figure 4.3: ROC Curves of Machine Learning Models for Predicting Vital Status in CRC Patients (Training Set)

Figure 4.3 presents the ROC curves of the five ML models applied to predict vital status in CRC patients based on the training set. These curves illustrate each model's sensitivity-specificity trade-off across thresholds, with the AUC representing their ability to differentiate between deceased and surviving patients.

Among all models, the XGBLinear achieved the highest AUC (0.96), indicating superior discriminative performance. The RF and svmRadial models also performed robustly, with AUCs of 0.89 and 0.88, respectively, suggesting strong generalization with balanced classification. The ANN model followed with an AUC of 0.85, maintaining reliable sensitivity and precision. The svmLinear algorithm achieved an AUC of 0.84, demonstrating its effectiveness in identifying mortality cases, albeit with a slightly more conservative curve.

Overall, all models exhibited high AUCs (>0.83), affirming their predictive value. Notably, XGBLinear, RF, and svmRadial demonstrated excellent performance for vital status prediction and are promising tools for risk stratification in clinical settings.

Feature selection results for vital status prediction are presented in Table 4.9. Age, surgery, and stage category were the only predictors selected by all five ML models, underscoring their consistent and critical relevance in CRC survival prediction. Gender, radiation, site category, and histology category were each retained by four models, further highlighting their strong predictive utility. Chemotherapy, palliative care, and regional lymph node involvement were selected by three models, while marital status and immunotherapy appeared in only two (ANN and xgbLinear). The variation in selected features—particularly among demographic and treatment-related variables—reflects the distinct ways in which each algorithm prioritizes information. Nonetheless, the overlap in core clinical predictors across models suggests a high degree of consensus on the most influential factors in determining vital status among CRC patients.

Table 4.9: Features Retained in Final Machine Learning Models for Vital Status Prediction

Variable	RF	ANN	svmLinear	svmRadial	XgbLinear
Gender	✓	✓	✓		✓
Age	✓	✓	✓	✓	✓
Marital		✓			✓
Surgery	✓	✓	✓	✓	✓
Radiation	✓	✓		✓	✓
Chemotherapy	✓			✓	✓
Immunotherapy		✓			✓
Palliative	✓	✓		✓	
Regional Lymph Node	✓	✓		✓	
Stage at Diagnosis	✓	✓	✓	✓	✓
Primary Tumor Site	✓	✓		✓	✓
Histology – Behavior	✓	✓		✓	✓

A check mark (✓) indicates that the corresponding variable was selected in the final version of each model. Blank cells denote features that were excluded during the embedded feature selection process.

4.4 Discussion

This chapter presented an in-depth comparison of five ML models, including Random Forest (RF) and Support Vector Machines using both linear and radial kernels (svmLinear and svmRadial), Extreme Gradient Boosting using a linear booster (XGBLinear), and Artificial Neural Networks (ANN)—for predicting three critical CRC outcomes: cancer status, recurrence, and vital status. Using clinical data from 1,314 patients treated at King Fahd

Specialist Hospital in Dammam, Saudi Arabia, the primary objective was to identify robust and generalizable models that could enhance prognostic accuracy, inform personalized treatment strategies, and support clinical decision-making in a regional healthcare context.

All models demonstrated acceptable to strong performance, achieving AUC scores above 0.78 in the majority of tasks, underscoring the feasibility of applying ML to CRC outcome prediction. For cancer status prediction, ANN and XGBLinear emerged as the most balanced and consistent classifiers, demonstrating strong generalization across training and validation sets. While RF offered the highest specificity and PPV, its relatively lower sensitivity suggests a conservative bias, making it well-suited for confirmatory assessments rather than broad screening. The SVM models revealed distinct trade-offs—svmRadial prioritized recall, while svmLinear emphasized specificity—highlighting their respective roles in context-dependent applications.

For recurrence prediction, XGBLinear again outperformed other models during both the training phase (AUC = 0.8172) and validation phase (AUC = 0.7641), showing the best overall balance between sensitivity and specificity. RF maintained the highest validation sensitivity (0.7717), indicating strong potential for capturing recurrence cases early, which is essential in clinical surveillance. ANN offered the highest training sensitivity (0.8611) but showed moderate performance on validation, suggesting some overfitting. SVM models maintained moderate and acceptable performance with different emphases—svmLinear favoring recall and svmRadial offering higher specificity—demonstrating their flexibility in clinical decision contexts depending on the cost of misclassification.

In the vital status prediction task, the XGBLinear model achieved the highest AUC (0.9563) on the training set but experienced a decline in validation (AUC = 0.7971), indicating a potential overfitting issue. RF, by contrast, demonstrated stable performance across datasets with a validation AUC of 0.8439 and balanced sensitivity (0.7276) and specificity (0.8431), making it the most reliable overall classifier for survival prediction. ANN also maintained strong performance (validation AUC = 0.8357), with balanced metrics that support its utility in mortality risk stratification. Both SVM variants achieved AUCs above 0.82 in training and 0.82–0.83 in validation, reinforcing their reliability as secondary models when precision or recall is prioritized.

Feature selection analyses revealed several consistently retained predictors across all three outcomes. Variables such as stage at diagnosis, primary tumor site, regional lymph node involvement, and surgery were selected by nearly all models, reaffirming their prognostic importance. Radiation, chemotherapy, and age were also frequently retained, particularly in recurrence and survival tasks. Notably, demographic features such as gender and marital status showed variable importance across models, suggesting that their predictive value may depend on model architecture or outcome-specific interactions.

This study is among the first in the region to apply a rigorous, nested cross-validation and bootstrap-based comparison of ML models for multi-outcome CRC prediction using hospital-based clinical data. Despite its strengths, several limitations merit consideration. First, the dataset, although well-structured and representative of tertiary care, is limited to a single institution, which may restrict external generalizability. Second, the low prevalence of recurrence constrained model calibration, even with class balancing strategies such as SMOTE and up-sampling. Third, the study used only structured clinical variables, excluding genomic, molecular, and behavioral data, which could enhance model precision.

In conclusion, the findings of this chapter underscore the promise of ML in predicting recurrence, cancer status, and vital status in CRC patients. While XGBLinear and RF consistently performed best across outcomes, ANN offered a strong alternative when sensitivity is clinically prioritized. The models developed here demonstrate potential to support individualized care pathways and improve CRC outcome monitoring in Saudi Arabia. Future work should focus on expanding data sources through multi-center collaborations, incorporating multi-omics data, and prospectively validating these models in clinical workflows to optimize their practical utility and impact.

5. The Impact of Transitions in Cancer Control Policies on Colorectal Cancer Prevalence, Incidence, and Mortality

5.1 Background

Throughout the previous few decades, the global burden of cancer has continued to rise, making it a major public health challenge worldwide [136]. In response, countries have implemented a range of cancer control initiatives, including the establishment of population-based cancer registries [137], national screening programs [138], and public health campaigns aimed at reducing Changeable risk factors like smoking, lack of exercise, and poor dietary habits [139]. These evolving strategies reflect a global commitment to improving cancer prevention, early detection, and access to timely, effective treatment. Tracking changes in cancer rates—including prevalence, incidence, and death—is essential for assessing how well these strategies are working [140]. For example, the introduction of screening programs often results in an initial rise in reported cases due to enhanced detection [141], while the decline in cancer-related death rates over the past 30 years has largely been attributed to improved screening and reduced tobacco use—outcomes driven by public health campaigns and legislative actions, as highlighted by the American Association for Cancer Research [142].

Evaluating the effectiveness of CRC control policies requires a comprehensive assessment of their impact on prevalence, incidence, and mortality rates, as well as their cost-effectiveness and population-level adherence. Recent literature underscores the multifaceted nature of CRC prevention, highlighting the importance of evidence-based policies that are comprehensive, accessible, and economically sustainable [143]. Effective CRC control strategies often include early detection initiatives, population-based screening programs, and efforts to ensure equitable access to healthcare services [144].

Studies from various regions have shown that multi-level prevention strategies can be effective but often vary in impact due to differences in implementation, awareness, and resources [145]. Cost-effectiveness analyses emphasize the benefits of early-stage detection in reducing treatment costs and improving outcomes [144]. Screening participation,

particularly through methods like the fecal immunochemical test (FIT), has been linked to significant reductions in mortality[146], while innovations like multi-target stool DNA tests also show promise [147]. Policy-level interventions, such as health insurance mandates, have shown measurable impact; for example, Preston et al. (2023)[148] found that such mandates increased up-to-date CRC screening by 2.8 percentage points post-ACA implementation, equivalent to an estimated 2.37 million additional screenings. Their findings underscore the role of reduced out-of-pocket costs in increasing access to preventive services and advancing precision public health. Despite these advancements, challenges remain in ensuring public adherence due to complex guidelines [149], underscoring the need for context-sensitive policies, clear communication, and continued efforts to reduce disparities in CRC care.

To support effective monitoring and timely policy interventions, control charts have become essential tools in analyzing cancer trends. Developed by Walter A. Shewhart in the 1920s as part of Statistical Process Control (SPC) [150], these charts are now widely used in healthcare to track indicators such as incidence, recovery times, and mortality rates [151]. Their role in cancer epidemiology has expanded, helping researchers detect unexpected variations or shifts in trends. Advanced models - like the Cumulative Detection Error Chart (CDEC) - are particularly effective for analyzing incomplete or irregular healthcare data [151]. A recent population-based study[152] highlighted that transitions between ICD coding systems can significantly affect disease prevalence estimates, emphasizing the need to account for such changes in longitudinal analyses. Another study demonstrated that while control charts are effective in monitoring chronic disease trends, their reliability depends heavily on data quality and methodological rigor [153].

Although global research has highlighted the value of cancer control policies and statistical tools like control charts, few studies have integrated population level policy evaluations with SPC techniques particularly in the context of CRC trends in KSA. Existing literature in the Kingdom has largely focused on descriptive statistics, registry-based summaries, or the evaluation of individual policies, such as the Saudi Cancer Registry or national CRC screening programs. However, a comprehensive analysis that links cancer control policy transitions to changes in CRC prevalence, incidence, and mortality through robust SPC methodologies remains largely unexplored.

This research seeks to address this gap by assessing how key policy interventions have influenced CRC patterns in KSA. Using longitudinal, population-based data from 1990 to 2021, this study applies control charts to detect statistically significant shifts in CRC prevalence, incidence, and mortality before and after key policy implementations. By aligning policy timelines with statistical trend monitoring, the study seeks to assess the effectiveness of KSA's cancer control strategies and provide evidence to inform data-driven healthcare planning and policy optimization.

5.2 Methods

5.2.1 Data source, exclusion, and inclusion approach

Data for this study were extracted from the GBD Report 2021, a comprehensive epidemiological database that provides systematically collected estimates for incidence, prevalence, and mortality across various diseases and risk factors. The GBD framework covers seven super-regions, 21 regions, and 204 countries and territories, utilizing rigorous statistical modeling and validation methodologies to ensure data accuracy and comparability across time and locations.

To acquire the data that is required, this study utilized the Global Health Data Exchange (GHDx) tool (<https://ghdx.healthdata.org/gbd-2021>, accessed on 01 February 2025). The selected cause was 'colon and rectum cancer', and the measures included Prevalence, Incidence and Deaths.

Additionally, population estimates were extracted separately as GBD population estimates, independent of disease-specific metrics.

For this study, data were specifically extracted for KSA, focusing on CRC trends. The dataset included 14 Age groups: 20–24, 25–29, 30–34, 35–39, 40–44, 45–49, 50–54, 55–59, 60–64, 65–69, 70–74, 75–79, 80–84, and 85 years and older, 4 Time intervals: 1990–1995, 1996–2007, 2008–2016, 2017–2021 and gender

The GBD Results Tool was used to extract only the relevant epidemiological metrics, ensuring that all non-selected variables were systematically excluded. This structured approach allows for a precise assessment of CRC burden trends in KSA, facilitating informed

analysis and policy recommendations. The study then selected the included nation based on the established inclusion criteria, ensuring consistency and relevance to the research objectives.

5.2.2 Defining Study Periods by Policy Transitions:

The evolution of cancer control policies in KSA reflects a gradual and strategic transformation, aligned with national healthcare priorities and global best practices. This development can be categorized into distinct phases, each characterized by specific advancements and challenges.

1. 1990–1995: The Foundation Phase

This phase represented the initial steps taken by KSA to confront the escalating impact of cancer. The focus was primarily on collecting data and building the foundational infrastructure for healthcare, with an emphasis on cancer incidence and prevalence. In 1992, the Saudi Cancer Registry (SCR) was established as a population-based registry[154]. Its primary aim was to document cancer cases and provide essential data for understanding the disease's trends across the Kingdom[155]. However, challenges such as incomplete data linkage with mortality statistics, primary care, and hospital records limited the effectiveness of these early efforts[156], [157]

The early 1990s also saw the introduction of palliative care services by King Faisal Specialist Hospital and Research Centre (KFSH&RC) [158]. Initially launched as a home care service, this initiative marked a significant step towards patient-centered cancer care in KSA[159].

The focus of this stage is to establish foundational data collection without structured programs or prevention strategies, which limit its effectiveness in comprehensive cancer control.

2. 1996–2007: Early Prevention and Care

This phase marked the beginning of proactive cancer care, with a focus on the formalization of palliative care and early awareness initiatives targeting modifiable risk factors. In 1996, KFSH&RC transitioned its palliative care services into a hospital-wide consultation program[160] formally establishing the Palliative Care Medicine section within the

Department of Oncology. This was followed by the founding of the Saudi Oncology Society (SOS) in 2007 to advance scientific knowledge, improve cancer management, and develop national guidelines for cancer care [161], [162]. In the same year, the Saudi Cancer Society, headquartered in Riyadh, and the Abdul Lateef Charitable Screening Center formed a non-governmental partnership to create the first breast cancer screening facility in the country [163]. This period laid the foundation for future structured prevention programs while addressing cancer care needs for advanced cases.

3. 2008–2016: Structured Screening and Prevention

This phase saw the scaling up of national-level programs and the introduction of comprehensive policies for cancer prevention and control. During this period, organized screening programs for various cancers were implemented, along with increased public health initiatives targeting lifestyle-related risk factors. A significant milestone in 2008 was the launch of the National Program for Cancer Prevention, which laid the groundwork for structured national interventions [164]. The Saudi Ministry of Health (MOH) initiated a localized, community-focused intervention known as the Crown Health Project in the Al-Jouf region [165], a northern region of KSA. This initiative aimed to assess the feasibility and effectiveness of targeted prevention strategies with the ultimate goal of expanding the program nationwide [166]. The project aimed to prevent non-communicable diseases through a multi-step approach that included screening for lifestyle-related risk factors, reviewing medical histories, and conducting physical and biochemical assessments.

A new national health care plan was established in KSA in 2011[167], with an emphasis on raising the standard of preventative, curative, and rehabilitative care. one of the primary goals. KSA has been enacting stringent legislation that forbids smoking indoors and in public spaces and controls the labeling and packaging of tobacco products. Additionally, the Ministry of Health launched a four-year (2008–2011) eHealth project [168] to enhance digital health services across healthcare facilities. By 2011, the strategy was fully integrated into hospitals in major cities, modernizing healthcare delivery and data management.

A significant advancement occurred in 2012 with the introduction of the Nationwide Breast Cancer Screening Program by the Ministry of Health [169]. This initiative reflected growing

public awareness and policy-driven action to enhance the timely identification and clinical management of breast cancer. Further strengthening cancer control efforts, in 2016, the Saudi National Cancer Institute (SANCI) was established under the Saudi Health Council [170] to coordinate and develop national strategies for cancer control.

This period marked the transition from awareness campaigns to structured programs, focusing on prevention, early detection, and risk factor management. Additionally, efforts were directed towards developing structured cancer screening programs and national guidelines for cancer management, with key contributions from the Saudi Oncology Society in establishing standardized, evidence-based protocols.

4. 2017–2021: Vision 2030 Reforms

This period was marked by transformative healthcare reforms aligned with Vision 2030, with a strong emphasis on preventive care and addressing modifiable risk factors. A key milestone in 2017 was the implementation of excise taxes, including a 100% tax on tobacco products and energy drinks and a 50% tax on soft drinks, as part of broader public health efforts to reduce tobacco use and encourage healthier lifestyles[167]. In the same year, CRC screening programs were integrated into national healthcare strategies, alongside the introduction of advanced therapies such as proton therapy at King Fahd Medical City[171]. The Al-Kharj CRC screening program was launched in 2017 as a five-year initiative (2017–2022) in a central urban area of Riyadh Province to evaluate participation and effectiveness in detecting CRC in average-risk individuals [172].

As of 2019, the revised Saudi National Immunization Program included HPV vaccination as a standard part of immunization for females to help prevent HPV-related cancers, while the vaccine remained optional for males in KSA[173].

Further reinforcing these efforts, Vision 2030's Health Sector Transformation Program placed a strong emphasis on non-communicable diseases (NCDs) and cancer prevention[174].

This phase underscored the importance of preventive care, leveraging public policy interventions to reduce the cancer burden while promoting healthier lifestyles through

taxation, regulation, and public awareness campaigns. The reforms aimed to enhance healthcare equity and accessibility, ensuring that individuals across KSA benefit from early detection and preventive healthcare measures.

5.2.3 Statistical analyses

Annual analyses were conducted throughout the GBD timeframe (1990–2021), utilizing descriptive metrics—including averages, standard deviations, and proportions—to outline the population’s demographic characteristics.

We fit counts of cases for CRC prevalence, incidence and deaths in each age group and sex stratum for the j^{th} year ($j=1, \dots, m$; $m=32$) to three negative binomial regression models (M1), a generalization of Poisson regression models (M2) that relax the restrictive assumption that the variance is equal to mean. To reduce Type 1 error risks, the likelihood ratio test was used to evaluate the model's goodness of fit, using a significance threshold of $\alpha = 0.01$. To determine the most appropriate model, the ratio of residual deviance to degrees of freedom (df) was examined [175] [176] [177] of the Poisson regression and negative binomial models were also examined.

The models incorporated an offset term defined as the natural logarithm of the population size for each combination of age and sex. Age group and sex were the main factors in the three models. Together with the annual regression coefficient $\hat{B}_j = (b_{kj})$, $k = 0, 1, \dots, (p - 1)$, the models generated annual estimates for age- and sex-adjusted prevalence, incidence, and mortality. Each model involved estimating a total of 15 parameters ($p = 15$). Female was used as the reference group for sex, while the 40–44 year age category served as the baseline for age. For each year j , the model parameters comprised an intercept term (b_{0j}) and slope coefficients ranging from $b_{1j}, \dots, b_{14j}$.

We calculated the mean annual percentage change in the age- and sex-standardized prevalence, incidence, and mortality of CRC for four time segments, Segment 1 (1990–1995), Segment 2 (1996–2007), Segment 3 (2008–2016) and Segment 4 (2017–2021). This segmentation allows for an assessment of trends in CRC burden over time, facilitating a better understanding of disease progression and the effectiveness of public health interventions.

For each year of the GBD research period (1990–2021), we tracked and signaled changes in the regression model parameter estimates using Hotelling's T2 control chart [178][179]. This methodology was used to monitor changes in the trends of CRC mortality, incidence, and prevalence over time. Our hypothesis was that consistency in the projected CRC burden would be shown by stability in the model parameters.

The regression model parameter estimations are simultaneously tracked using Hotelling's T2 statistic; if the statistic rises above a predetermined control limit value[180], an out-of-control signal is generated [181]. This method makes it possible to spot notable variations in CRC prevalence, incidence, and death, which could be the result of underlying changes brought about by policy modifications, public health initiatives, or other variables affecting disease patterns.

In order to identify the existence of autocorrelation[182] (and calculate the autocorrelation coefficient) between years, we employed the Durbin- Watson[183] test statistics, since prevalence, incidence and mortality estimates for later years can include the same people. Autocorrelation significantly impairs control chart performance[184] by violating the assumption of independence among observations, leading to increased false alarms and reduced detection capabilities[185] . show that ignoring autocorrelation leads to frequent false signals, undermining the effectiveness of SPC systems[186]

In our study, we employed a U-statistic approach to mitigate the effects of autocorrelation in assessing changes in regression model parameters over time[187]. U-statistics, introduced by Wassily Hoeffding in 1948[188], are a class of unbiased estimators defined as the average value of a real-valued function computed over all possible subsets of a given size from a sample. This methodology has been widely applied in various fields[187].

The coefficients for the negative binomial regression model were estimated as $\hat{\beta}_j = (b_{kj})$ for $k = 0, 1, \dots, (p - 1)$, where p represents the total number of estimated parameters. These coefficients were obtained through MLE. To account for autocorrelation, the estimated parameters were adjusted using a U-statistic [189][187]. defined as

$$\mathbf{U}_j = (b'_{0j}, b'_{1j}, \dots, b'_{pj})' = \begin{cases} \hat{\mathbf{A}}_1, & \text{when } j = 1 \\ \hat{\mathbf{A}}_j - \boldsymbol{\Sigma}_{j,j-1} \boldsymbol{\Sigma}_{j-1,j-1}^{-1} \hat{\mathbf{A}}_{j-1}, & \text{when } j = 2, \dots, m \end{cases}$$

Here b'_{0j} represents the adjusted estimated intercept, while $b'_{1j}, \dots, b'_{pj}$ correspond to the adjusted estimated slope parameters for the j^{th} year, The adjusted parameter vector is defined as $\hat{\mathbf{A}}_j = \hat{\boldsymbol{\beta}}_j / (1 - \varphi)$, where φ denotes the estimated value of the autocorrelation parameter, computed using the Durbin-Watson test statistic. Additionally, $\boldsymbol{\Sigma}_{(.,.)}$ denotes the covariance matrix of $\hat{\mathbf{A}}_j$. The vectors $\mathbf{U}_j$ s are assumed to be independently and normally distributed, with their respective mean vector and covariance matrix expressed as follows:

$$\boldsymbol{\mu}_{U_j} = \begin{cases} \hat{\mathbf{A}}_1, & \text{when } j = 1 \\ \boldsymbol{\mu}_{\hat{\mathbf{A}}_j} - \boldsymbol{\Sigma}_{j,j-1} \boldsymbol{\Sigma}_{j-1,j-1}^{-1} \boldsymbol{\mu}_{\hat{\mathbf{A}}_{j-1}}, & \text{when } j = 2, \dots, m \end{cases}$$

$$\boldsymbol{\Sigma}_{U_j} = \begin{cases} \boldsymbol{\Sigma}_{1,1}, & \text{when } j = 1 \\ \boldsymbol{\Sigma}_{j,j} - \boldsymbol{\Sigma}_{j,j-1} \boldsymbol{\Sigma}_{j-1,j-1}^{-1} \boldsymbol{\Sigma}_{j-1,j}, & \text{when } j = 2, \dots, m \end{cases}$$

In our chapter, U statistics are specifically employed to monitor the estimated regression model parameters for CRC during a 32-year period (1990-2021). This approach allows us to detect significant changes in prevalence, incidence and mortality estimates when transitioning between cancer control policies.

The use of U statistics in this study is complemented by control charts, which help visualize the stability of the estimates and identify any out-of-control observations that may indicate significant shifts in CRC prevalence, incidence and mortality. where the estimate for the j th year is adjusted for correlation in the preceding (i.e., $j - 1$) year [190][187].

This is what Hotelling's T2 statistic[180] [191][187]for the j th year is:

$$T_j^2 = (\mathbf{U}_j - \boldsymbol{\mu}_U)' \boldsymbol{\Sigma}_U^{-1} (\mathbf{U}_j - \boldsymbol{\mu}_U)$$

and plotted against an upper control limit (UCL) of.

$$UCL = \frac{p(m-1)(m+1)}{m(m-p)} F_{\alpha,p,m-p}$$

and zero as the lower control limit (LCL). With mean $\boldsymbol{\mu}_U$ and covariance matrix $\boldsymbol{\Sigma}_U$, $\mathbf{U}_j$ is the vector of adjusted regression model parameters, which are presumed to be independently and normally distributed.

In the previous equation, α is the nominal level of significance, and F is the critical value of the F distribution with degrees of freedom p and $m-p$. When $\alpha = (1/370)$, The choice of α aligns with standard SPC practices, where the expected Average Run Length (ARL) for an in-control process is conventionally set to 370 observations before a false alarm occurs [192]. The control limit was 24.2. One can estimate the covariance matrix Σ_U and mean vector μ_U from a training dataset or reference sample[152].

Finding the mean and covariance of U_j across all j allows one to estimate the mean vector and covariance matrix from U_j in the absence of a reference sample [152]. When test statistics fall within the LCL and UCL, it is considered an in-control observation; otherwise, it is considered an out-of-control observation. To determine if cancer control policies had an impact on the prevalence, incidence and mortality estimates of a CRC, at least one out-of-control observation during a transition period was required.

This study monitors the entire transition period to capture multiple policy changes that occurred at different points within the period. In this study, the transition periods correspond to major shifts in healthcare policies and screening programs. The first transition period was set between 1995 and 1999, coinciding with the early efforts in structured cancer control. The second transition period spanned 2006 to 2010, marking the introduction of national cancer screening initiatives, and the third transition period covered 2015 to 2019, reflecting the implementation of Vision 2030 reforms that emphasized preventive healthcare and CRC screening. By analyzing the entire transition period, this approach ensures that all relevant policy shifts within each period are accounted for, providing a comprehensive assessment of their impact on CRC prevalence and healthcare trends.

5.3 Results

5.3.1 Characteristics of the Study Population

Table 5.1 provides a description of the research population's demographics for a few chosen study years. The study population aged 20 years and older increased from 7.6 million in 1990 to 27.6 million in 2021. During this period, the average age rose slightly from 36.8 years (SD: 13.7) to 38.7 years (SD: 12.2). The proportion of females remained relatively stable, ranging from 39.4% in 1990 to 39.7% in 2021.

Table 5.1: Demographic Characteristics of the Study Population in Selected Years

Characteristics	1990	1995	2000	2005	2010	2015	2021
N	7645015	9442757	11138698	13565253	17694075	22430194	27570815
Age, mean (SD)	36.8 (13.7)	37.1 (13.5)	37.4 (13.5)	37.2 (13.2)	36.8 (12.7)	37.2 (12.3)	38.7 (12.2)
Female (%)	39.4	39.5	40.5	41.0	40.6	40.0	39.7

5.3.2 Goodness of fit tests for regression models

We assessed the goodness of fit for M1 (Negative Binomial model) and M2 (Poisson model) across models constructed using prevalence, incidence, and mortality data. The likelihood ratio test rejected the null hypothesis (dispersion parameter = 0) for all three models ($p < 0.01$), indicating significant over-dispersion in the data.

Additionally, the ratio of residual deviance to degrees of freedom (df) was consistently lower for the Negative Binomial model (M1) than for the Poisson model (M2), confirming its superior fit. The Poisson model performed particularly poorly for the prevalence dataset, where the ratio was 91.51, compared to 12.90 for the Negative Binomial model, highlighting substantial over-dispersion. Similarly, for the incidence dataset, the Poisson model had a ratio of 18.71, whereas the Negative Binomial model had a lower value of 11.47.

For the mortality dataset, while both models produced similar ratio values, the likelihood ratio test yielded a p-value of 0, confirming the presence of over-dispersion. This result indicates that the Negative Binomial model provides a statistically better fit even for the mortality dataset.

Given these findings, we selected the Negative Binomial model as the preferred model for all three datasets to ensure consistency in analysis and better account for potential over-dispersion in the data.

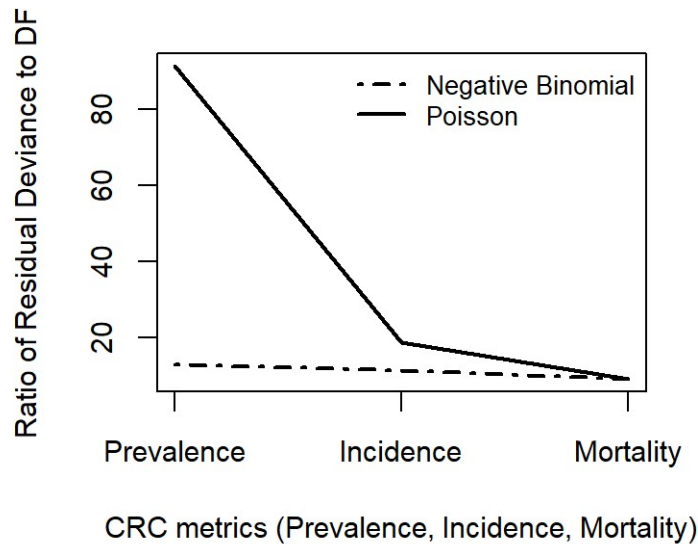


Figure 5.1: Comparing Model Fit for Poisson and Negative Binomial Regression for prevalence, incidence and mortality.

5.3.3 Age- and sex-adjusted prevalence, incidence and mortality

Table 5.2 presents the prevalence, incidence, and mortality rates (per 100,000 population) along with the average annual rate of change (%) across the four key policy transition periods of cancer control efforts in KSA. These periods reflect major milestones in prevention, screening, and treatment policies, shaping the overall burden of CRC.

During the Foundation Phase (1990–1995), the prevalence rate increased from 27.2 per 100,000 in 1990 to 32.7 per 100,000 in 1995, marking the initial phase of cancer data collection with limited structured prevention efforts. Similarly, the incidence rate increased from 5.7 to 6.4 per 100,000, while the mortality rate showed minimal growth, rising from 4.3 to 4.5 per 100,000 during this period. The average annual rate of change in prevalence was 8.3% (95% CI: 8.02–8.61), while incidence and mortality increased at annual rates of 6.91% (95% CI: 6.26–7.55) and 5.56% (95% CI: 4.82–6.31), respectively.

The Early Prevention and Care Phase (1996–2007) saw structured cancer control efforts, including the establishment of screening initiatives and national guidelines. The prevalence rate rose from 33.7 per 100,000 in 1996 to 53.2 per 100,000 in 2007, with an average annual

increase of 8.25% (95% CI: 8.15–8.35). Incidence and mortality also followed an upward trend, reaching 10.2 and 5.9 per 100,000, respectively, with annual increases of 8.44% (95% CI: 8.27–8.60) and 6.41% (95% CI: 6.21–6.61), respectively.

In the Structured Screening and Prevention Phase (2008–2016), national screening programs and prevention initiatives were introduced, leading to an increase in prevalence from 55.1 per 100,000 in 2008 to 67.6 per 100,000 in 2016, although at a slightly lower annual growth rate of 7.27% (95% CI: 7.09–7.46). Incidence and mortality continued to rise, reaching 11.4 and 5.2 per 100,000, respectively, with annual increases of 6.23% (95% CI: 6.05–6.42) and 3.44% (95% CI: 3.21–3.67), respectively.

The Vision 2030 Reform Phase (2017–2021) emphasized preventive care, taxation policies, and national CRC screening programs. The prevalence rate reached 77.1 per 100,000 in 2021, but the rate of increase slowed to 4.99% (95% CI: 4.54–5.43), suggesting a potential stabilization in disease burden. Incidence rose to 12.5 per 100,000, while mortality increased at the slowest rate among all periods, reaching 5.3 per 100,000, with an annual increase of 3.85% (95% CI: 3.36–4.34).

Across all policy transition periods, prevalence, incidence, and mortality exhibited an overall upward trend, albeit at varying rates depending on policy efforts. The most significant growth was observed in the Early Prevention and Care Phase (1996–2007), while a gradual stabilization occurred in the most recent period (2017–2021), reflecting the impact of cancer prevention and early detection strategies in KSA.

Table 5. 2: Prevalence, incidence and mortality and their average annual rate of change (%)

CRC Metric	Segment 1: 1990—1995			Segment 2: 1996—2007			Segment 3: 2008—2016			Segment 3: 2017—2021		
	1990	1995	Average annual rate of change (95% CI)	1996	2007	Average annual rate of change (95% CI)	2008	2016	Average annual rate of change (95% CI)	2017	2021	Average annual rate of change (95% CI)
Prevalence	27.2	32.7	8.3 (8.02, 8.61)	33.7	53.2	8.25 (8.15, 8.35)	55.1	67.6	7.27(7.09 , 7.46)	69.4	77.1	4.99 (4.54 , 5.43)
Incidence	5.7	6.4	6.91 (6.26, 7.55)	6.5	10.2	8.44 (8.27, 8.60)	10.3	11.4	6.23(6.05, 6.42)	11.6	12.5	5.12(4.71, 5.54)
Mortality	4.3	4.5	5.56(4.82, 6.31)	4.6	5.9	6.41(6.21 , 6.61)	5.8	5.2	3.44 (3.21, 3.67)	5.1	5.3	3.85(3.36, 4.34)

Note: Age- and sex-adjusted prevalence per 100,000 population

5.3.4 Result

The control charts for CRC prevalence, incidence, and mortality highlight significant deviations in specific years across the four policy transition phases in KSA. These deviations suggest shifts in CRC trends, likely influenced by major healthcare reforms, screening initiatives, and changes in reporting practices.

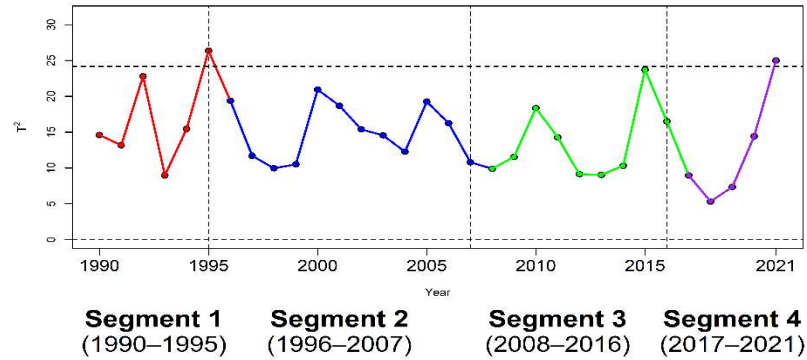
In the prevalence control chart, CRC prevalence followed a consistent increasing trend across all segments. During the first segment (1990–1995), the year 1995 exceeded the upper control limit (UCL), indicating a statistically significant deviation. This shift likely corresponds with the establishment of the Saudi Cancer Registry (SCR) in 1992, which improved cancer case documentation and led to more complete case identification. In segment two (1996–2007) and segment three (2008–2016), prevalence remained within control limits, reflecting a stable trend as screening and prevention programs became more structured. However, in segment four (2017–2021), the year 2021 again exceeded the UCL, suggesting a notable deviation in CRC prevalence. This could be attributed to expanded national CRC screening programs, which enhanced case detection. Alternatively, it may reflect delayed diagnoses due to COVID-19-related healthcare disruptions, resulting in an increased number of cases being diagnosed in 2021.

The incidence control chart follows a similar trajectory. The year 1995 in segment one (1990–1995) exceeded the UCL, aligning with improved cancer reporting efforts initiated by the SCR. In segment two (1996–2007) and segment three (2008–2016), incidence rates fluctuated but remained within control limits, implying a gradual and expected increase in newly diagnosed cases. However, in segment four (2017–2021), the year 2021 approached the UCL, with a T^2 value of 24.13 against the control limit of 24.2, indicating a marginal but notable deviation in CRC incidence. This increase is likely due to the expansion of CRC screening programs and heightened awareness, leading to higher detection rates. Additionally, international studies [193] have reported delays in routine screenings during the COVID-19 pandemic, leading to a backlog of undiagnosed cases that may have contributed to the rise in CRC diagnoses in 2021. While no specific study in KSA has examined this impact, the observed increase in incidence aligns with global trends where cancer detection rates rebounded following disruptions in healthcare services during the

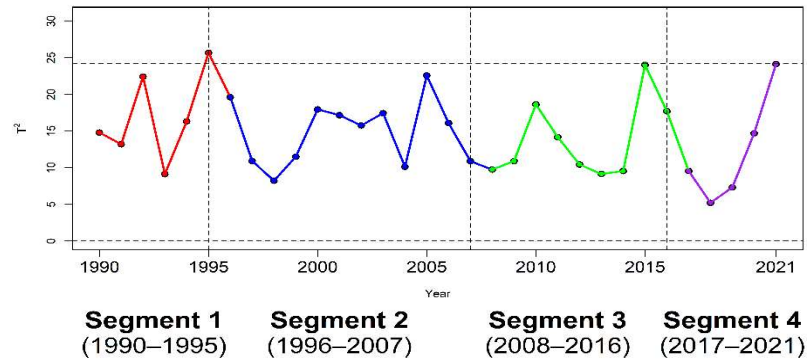
pandemic. Further investigation is warranted to determine the extent to which these factors influenced CRC incidence trends in KSA.

The mortality control chart presents a somewhat different trend. The year 1992 exceeded the UCL in segment one (1990–1995), suggesting a significant shift in reported CRC-related deaths, potentially due to enhanced case tracking following the establishment of the SCR. In segment two (1996–2007) and segment three (2008–2016), no years exceeded the control limits, indicating relatively stable mortality trends. However, in segment four (2017–2021), the year 2021 again surpassed the UCL, signaling a deviation in CRC mortality. While this does not directly indicate an increase in late-stage diagnoses, it suggests that despite improved screening efforts, CRC-related deaths did not decline significantly. One possible explanation is that pandemic-related healthcare disruptions in 2020 may have delayed cancer diagnoses and treatments, leading to an increase in CRC-related mortality in 2021. Global studies have highlighted how delays in cancer screenings, postponed treatments, and healthcare system strain during the COVID-19 pandemic contributed to increased mortality across various cancer types [194][195]. The observed mortality shift in 2021 may align with these broader patterns. This underscores the need for continued improvements in early detection, treatment accessibility, and survivorship care to further reduce CRC mortality.

CRC Prevalence Control Chart



CRC Incidence Control Chart



CRC Mortality Control Chart

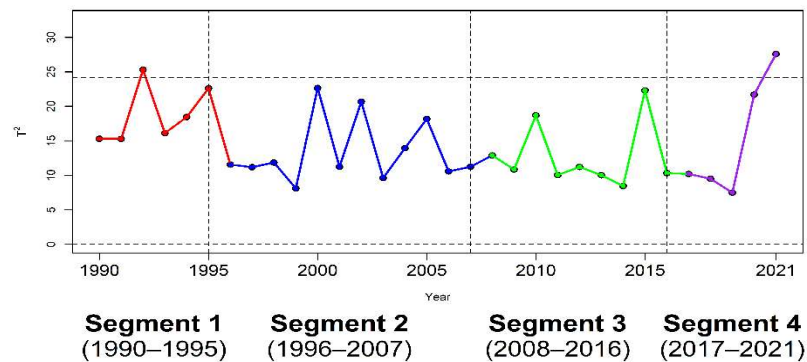


Figure 5.2: Control Chart for Prevalence, Incidence and Mortality; Horizontal dashed line represents the upper control limit of 24.2; Vertical dashed lines represent the end of the transition periods.

5.4 Discussion

We used control charts to monitor the estimated regression model parameters for CRC prevalence, incidence, and mortality in KSA over a 32-year period (1990–2021). This period was characterized by four major phases of cancer control policies, reflecting the evolving landscape of cancer prevention, screening, and treatment efforts. Unlike previous studies that assessed the impact of changes in disease classification systems[187], [196] (e.g., ICD transitions) on prevalence estimates, our study focused on identifying significant shifts in CRC trends corresponding to key policy and healthcare system transformations.

Our results revealed significant deviations in CRC prevalence and incidence in 1995, and in mortality during 1992. In 2021, all three outcomes showed elevated T^2 values. While prevalence and mortality exceeded the control limit, incidence approached it closely, with a T^2 value of 24.13 against a limit of 24.2, suggesting a borderline deviation. The 1995 deviation in prevalence and incidence coincides with the establishment of the Saudi Cancer Registry (SCR) in 1992, a milestone that improved cancer reporting and case documentation across healthcare facilities. The increased prevalence and incidence observed in this period likely reflect enhanced case identification rather than an actual surge in CRC occurrence. In contrast, the 2021 deviations align with the expansion of national CRC screening programs, which facilitated earlier detection of CRC cases, contributing to a recorded increase in prevalence and incidence. Additionally, disruptions in healthcare services during the COVID-19 pandemic may have led to delayed diagnoses in 2020, resulting in a backlog of cases being detected in 2021.

The mortality trends differed slightly from prevalence and incidence. The earliest deviation in mortality appeared in 1992, suggesting a possible increase in reported CRC-related deaths as an early impact of improved case tracking after the implementation of the SCR. CRC mortality trends remained relatively stable throughout much of the study period, with two notable deviations in 1992 and 2021. Despite improvements in screening and treatment accessibility, there was no consistent downward trend in mortality, highlighting ongoing challenges in treatment outcomes and survivorship care. The 2021 deviation in mortality may be attributed to multiple factors. While increased screening programs have improved early CRC detection, survival outcomes have not necessarily kept pace, indicating a gap in

treatment efficacy or accessibility. Additionally, pandemic-related disruptions in CRC care may have delayed diagnoses and treatments, potentially contributing to increased mortality in 2021. This finding underscores the need for further enhancements in early detection strategies, treatment accessibility, and survivorship care to reduce CRC-related mortality.

A key strength of our study is the application of control charts to assess shifts in CRC trends in response to policy interventions, providing a data-driven approach for evaluating the impact of healthcare reforms. However, our findings should be interpreted in the context of certain limitations. While control charts effectively identify statistical deviations, they do not pinpoint the specific underlying causes of these shifts. Additional analyses, incorporating stage-specific CRC data and treatment outcome assessments, would be necessary to determine whether the observed trends are linked to changes in early detection rates, treatment efficacy, or disease progression patterns.

Overall, our study highlights the positive impact of structured cancer control policies on improving CRC detection in KSA, particularly following the establishment of the national cancer registry and the expansion of screening programs. However, the continued challenges observed in CRC mortality trends suggest the need for integrated strategies focusing on both early detection and effective treatment interventions. Future research should explore longitudinal survival analyses, treatment accessibility disparities, and the role of lifestyle factors in shaping CRC trends in KSA.

6. Conclusion and Prospects for Future Research

6.1 Conclusion

This thesis investigated CRC outcomes in Saudi Arabia through a comprehensive and multi-tiered analytical framework. By integrating clinical data from 1,314 patients at King Fahd Specialist Hospital with population-level statistics from the Global Burden of Disease (GBD) 2021 dataset, the study addressed three major research objectives: (1) identifying significant clinical predictors of recurrence, survival, and cancer status; (2) evaluating the predictive performance of traditional statistical and ML models; and (3) assessing the impact of national cancer control policy transitions on CRC trends from 1990 to 2021.

Multivariate logistic and Cox regression analyses revealed that tumor site, disease stage, regional lymph node involvement, and treatment modalities (e.g., surgery, chemotherapy, immunotherapy) were significant predictors across all three CRC outcomes: recurrence, cancer status, and vital status. Age also emerged as a consistent demographic predictor. These findings confirm the utility of conventional regression models for outcome prediction and underscore the importance of clinical and pathological variables in patient stratification.

ML techniques further advanced the analytical framework. Five algorithms—RF, svmLinear, svmRadial, XGBLinear and ANN—were compared using a rigorous nested cross-validation design combined with bootstrap-based evaluation. ANN and XGBLinear models demonstrated the most balanced and generalizable performance for predicting cancer status, while XGBLinear and RF achieved the highest accuracy and AUC in recurrence prediction. For vital status, RF and ANN showed the most stable validation metrics, whereas XGBLinear, despite excelling in training, showed signs of overfitting. Feature selection consistently identified clinical predictors such as stage, site, treatment type, and age—affirming their central prognostic value and reinforcing the validity of embedded ML pipelines for CRC risk prediction.

At the population level, this study applied negative binomial regression and Hotelling's T^2 control charts to examine CRC prevalence, incidence, and mortality trends in KSA from

1990 to 2021. The analysis identified significant deviations in CRC prevalence in 1995 and 2021, in mortality in 1992 and 2021, and a near-threshold deviation in incidence in 2021. These shifts align with major health system milestones, including the establishment of the Saudi Cancer Registry and the expansion of national CRC screening programs. The findings demonstrate the measurable impact of policy interventions on CRC detection and reporting. However, the continued challenges in reducing mortality suggest persistent gaps in treatment accessibility or effectiveness, despite progress in early diagnosis.

Together, these results provide a unified framework that bridges patient-level prediction and policy-level evaluation. By integrating advanced analytics with real-world clinical and epidemiological data, the study offers actionable insights for clinicians, policymakers, and data scientists working to improve cancer outcomes. The work highlights the need for ongoing efforts to refine predictive tools, broaden data integration, and evaluate the full continuum of care—from diagnosis and treatment to long-term survivorship.

6.2 Prospects for Future Research

Despite the strengths of this thesis, several avenues remain for expanding and refining future research.

First, the clinical dataset used here was limited to a single tertiary care center, which may restrict generalizability. Expanding to multi-center or national registries would increase sample diversity, reduce selection bias, and enable more representative model training and validation. External validation of the developed models across different healthcare settings is a crucial next step.

Second, incorporating multi-modal data—including genetic markers, laboratory biomarkers, imaging data, and patient-reported outcomes—could improve predictive accuracy and reveal latent patterns not captured by structured clinical variables alone. As digital health infrastructure in Saudi Arabia continues to evolve, the integration of such data will become increasingly feasible.

Third, while five widely used ML algorithms were compared, future work could explore the application of deep learning architectures, ensemble stacking methods, and transformer-based models to better handle rare outcomes such as recurrence. Additionally,

model explainability should be prioritized using interpretable ML techniques (e.g., SHAP, LIME) to support clinician trust and real-world applicability.

From a policy evaluation perspective, further research could build on the time-series modeling and control chart analysis presented here by incorporating more granular data—such as stage-specific incidence, screening uptake, and survival by treatment modality. Incorporating forecasting methods like ARIMA, Prophet, or Bayesian structural time series models could also help simulate long-term outcomes of existing and proposed policy interventions.

Lastly, future studies should extend beyond traditional outcomes to include patient-centered metrics such as quality of life, psychosocial impact, treatment adherence, and care satisfaction. These dimensions are critical to understanding the full burden of CRC and aligning health services with the principles of equity and value-based care envisioned by Saudi Arabia’s Vision 2030 Health Sector Transformation Program.

In conclusion, this thesis has laid the foundation for data-driven CRC prediction and policy evaluation in Saudi Arabia. Continued interdisciplinary research—spanning clinical oncology, epidemiology, data science, and health policy—will be essential for scaling these insights and advancing personalized, equitable, and effective cancer control across the Kingdom and beyond.

REFERENCES

- [1] World Health Organization, “Noncommunicable diseases.” Accessed: Nov. 01, 2024. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases>
- [2] Ministry of Health, Kingdom of Saudi Arabia, World Health Organization, and United Nations Development Programme, “The case for investment in prevention and control of non-communicable diseases in the Kingdom of Saudi Arabia,” 2023. Accessed: May 15, 2025. [Online]. Available: <https://www.who.int/publications/i/item/9789240076992>
- [3] N. Saquib *et al.*, “Chronic disease prevalence among elderly Saudi men,” *Int J Health Sci (Qassim)*, vol. 11, no. 5, pp. 11–16, Dec. 2017.
- [4] M. S. Alzahrani, Y. S. Alharthi, J. K. Aljamal, A. A. Alarfaj, V. Vennu, and M. D. Noweir, “National and Regional Rates of Chronic Diseases and All-Cause Mortality in Saudi Arabia—Analysis of the 2018 Household Health Survey Data,” *Int J Environ Res Public Health*, vol. 20, no. 7, Apr. 2023, doi: 10.3390/ijerph20075254.
- [5] International Agency for Research on Cancer, “GLOBOCAN 2022: Saudi Arabia fact sheet,” 2024. Accessed: Jan. 15, 2025. [Online]. Available: <https://gco.iarc.who.int/media/globocan/factsheets/populations/682-saudi-arabia-fact-sheet.pdf>
- [6] M. Alyabsi, A. Alhumaid, H. Allah-Bakhsh, M. Alkelya, and M. A. Aziz, “Colorectal cancer in Saudi Arabia as the proof-of-principle model for implementing strategies of predictive, preventive, and personalized medicine in healthcare,” Mar. 01, 2020, *Springer*. doi: 10.1007/s13167-019-00186-x.
- [7] Ministry Of Health KSA, “Obesity.” Accessed: Nov. 01, 2024. [Online]. Available: <https://www.moh.gov.sa/en/awarenessplatform/ChronicDisease/Pages/Obesity.aspx>
- [8] M. Z. Aljulifi, “Prevalence and reasons of increased type 2 diabetes in Gulf Cooperation Council Countries,” May 01, 2021, *Saudi Arabian Armed Forces Hospital*. doi: 10.15537/SMJ.2021.42.5.20200676.
- [9] A. J. Al-Quwaidhi, M. S. Pearce, E. Sobngwi, J. A. Critchley, and M. O’Flaherty, “Comparison of type 2 diabetes prevalence estimates in Saudi Arabia from a validated Markov model against the International Diabetes Federation and other

- modelling studies,” *Diabetes Res Clin Pract*, vol. 103, no. 3, pp. 496–503, 2014, doi: 10.1016/j.diabres.2013.12.036.
- [10] M. Alqahtani *et al.*, “Risk factors of colorectal cancer among Saudi Population: Case-control study,” *J Family Med Prim Care*, vol. 9, no. 9, p. 5035, 2020, doi: 10.4103/jfmpe.jfmpe_895_20.
 - [11] A. M. Basudan, A. M. Basuwidan, M. Abudawood, R. Farzan, and M. A. Alfihili, “Comprehensive Retrospective Analysis of Colorectal Cancer Incidence Patterns in Saudi Arabia,” *Life*, vol. 13, no. 11, Nov. 2023, doi: 10.3390/life13112198.
 - [12] N. Alsadhan, S. A. Alhurishi, M. Pujades-Rodriguez, F. Shuweihdi, C. Brennan, and R. M. West, “Demographic and clinical characteristics associated with advanced stage colorectal cancer: a registry-based cohort study in Saudi Arabia,” *BMC Cancer*, vol. 24, no. 1, Dec. 2024, doi: 10.1186/s12885-024-12270-1.
 - [13] A. M. Alessa and A. S. Khan, “Epidemiology of Colorectal Cancer in Saudi Arabia: A Review,” *Cureus*, Jul. 2024, doi: 10.7759/cureus.64564.
 - [14] M. Alyabsi, F. Sabatin, M. Ramadan, and A. R. Jazieh, “Colorectal cancer survival among Ministry of National Guard-Health Affairs (MNG-HA) population 2009–2017: retrospective study,” *BMC Cancer*, vol. 21, no. 1, Dec. 2021, doi: 10.1186/s12885-021-08705-8.
 - [15] N. E. Elwali *et al.*, “Colorectal Cancer in Saudi Arabia: The Way Forward,” *Asian Pacific Journal of Cancer Prevention*, vol. 24, no. 1, pp. 13–19, 2023, doi: 10.31557/APJCP.2023.24.1.13.
 - [16] I. El Badisy *et al.*, “Risk factors affecting patients survival with colorectal cancer in Morocco: survival analysis using an interpretable machine learning approach,” *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-51304-3.
 - [17] World Health Organization, “Colorectal cancer”, Accessed: Nov. 01, 2024. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/colorectal-cancer>
 - [18] H. A. Rahman, M. A. Ottom, and I. Dinov, “Machine Learning-based Colorectal Cancer Prediction using Global Dietary Data,” Sep. 16, 2022. doi: 10.21203/rs.3.rs-2031672/v1.
 - [19] H. Qiu, S. Ding, J. Liu, L. Wang, and X. Wang, “Applications of Artificial Intelligence in Screening, Diagnosis, Treatment, and Prognosis of Colorectal Cancer,” *Current Oncology*, vol. 29, no. 3, pp. 1773–1795, Mar. 2022, doi: 10.3390/curroncol29030146.

- [20] O. Kennion, S. Maitland, and R. Brady, “Machine learning as a new horizon for colorectal cancer risk prediction? A systematic review,” *Health Sciences Review*, vol. 4, p. 100041, Sep. 2022, doi: 10.1016/j.hsr.2022.100041.
- [21] S. Alinia, M. Asghari-Jafarabadi, L. Mahmoudi, G. Roshanaei, and M. Safari, “Predicting mortality and recurrence in colorectal cancer: Comparative assessment of predictive models,” *Heliyon*, vol. 10, no. 6, Mar. 2024, doi: 10.1016/j.heliyon.2024.e27854.
- [22] Y. Liu, W. Du, Y. Guo, Z. Tian, and W. Shen, “Identification of high-risk factors for recurrence of colon cancer following complete mesocolic excision: An 8-year retrospective study,” *PLoS One*, vol. 18, no. 8 August, Aug. 2023, doi: 10.1371/journal.pone.0289621.
- [23] A. Cakmak *et al.*, “Predicting the predisposition to colorectal cancer based on SNP profiles of immune phenotypes using supervised learning models,” *Med Biol Eng Comput*, vol. 61, no. 1, pp. 243–258, Jan. 2023, doi: 10.1007/s11517-022-02707-9.
- [24] R. Talebi, C. A. Celis-Morales, A. Akbari, A. Talebi, N. Borumandnia, and M. A. Pourhoseingholi, “Machine learning-based classifiers to predict metastasis in colorectal cancer patients,” *Front Artif Intell*, vol. 7, 2024, doi: 10.3389/frai.2024.1285037.
- [25] B. Burnett *et al.*, “Machine Learning in Colorectal Cancer Risk Prediction from Routinely Collected Data: A Review,” Jan. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/diagnostics13020301.
- [26] H. Li *et al.*, “Colorectal Cancer Detected by Machine Learning Models Using Conventional Laboratory Test Data,” *Technol Cancer Res Treat*, vol. 20, 2021, doi: 10.1177/15330338211058352.
- [27] C. H. Chen, M. K. Tsai, C. Wen, and C. P. Wen, “A user-friendly objective prediction model in predicting colorectal cancer based on 234 044 Asian adults in a prospective cohort,” *ESMO Open*, vol. 6, no. 6, Dec. 2021, doi: 10.1016/j.esmoop.2021.100288.
- [28] A. Tamm *et al.*, “Establishing a colorectal cancer research database from routinely collected health data: the process and potential from a pilot study,” *BMJ Health Care Inform*, vol. 29, no. 1, Jun. 2022, doi: 10.1136/bmjhci-2021-100535.
- [29] P. C. Chen *et al.*, “A Prediction Model for Tumor Recurrence in Stage II–III Colorectal Cancer Patients: From a Machine Learning Model to Genomic Profiling,” *Biomedicines*, vol. 10, no. 2, Feb. 2022, doi: 10.3390/biomedicines10020340.

- [30] Y. Xu, L. Ju, J. Tong, C. M. Zhou, and J. J. Yang, "Machine Learning Algorithms for Predicting the Recurrence of Stage IV Colorectal Cancer After Tumor Resection," *Sci Rep*, vol. 10, no. 1, Dec. 2020, doi: 10.1038/s41598-020-59115-y.
- [31] M. Javaid, A. Haleem, R. Pratap Singh, R. Suman, and S. Rab, "Significance of machine learning in healthcare: Features, pillars and applications," *International Journal of Intelligent Networks*, vol. 3, pp. 58–73, Jan. 2022, doi: 10.1016/j.ijin.2022.05.002.
- [32] M. Liu *et al.*, "Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques."
- [33] Gareth. James, Daniela. Witten, Trevor. Hastie, and Robert. Tibshirani, *An introduction to statistical learning : with applications in R*. Springer : Springer Science+Business Media, 2017.
- [34] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
- [35] A. Esteva *et al.*, "A guide to deep learning in healthcare," Jan. 01, 2019, *Nature Publishing Group*. doi: 10.1038/s41591-018-0316-z.
- [36] X. Zhu and A. B. Goldberg, *Introduction to Semi-Supervised Learning*. 2009.
- [37] R. S. Sutton and A. G. Barto, "Reinforcement Learning: An Introduction Second edition, in progress."
- [38] O. Theobald, "Machine Learning For Absolute Beginners," 2017.
- [39] A. Amidi and S. Amidi, "CS 229-Machine Learning Super VIP Cheatsheet: Machine Learning," 2018. [Online]. Available: <https://stanford.edu/~shervine>
- [40] "Logistic Regression 12.1 Modeling Conditional Probabilities," 2013.
- [41] D. Liu, T. Li, and D. Liang, "Incorporating logistic regression to decision-theoretic rough sets for classifications," *International Journal of Approximate Reasoning*, vol. 55, no. 1 PART 2, pp. 197–210, 2014, doi: 10.1016/j.ijar.2013.02.013.
- [42] M. A. Arshad, S. Shahriar, and K. Anjum, "The Power Of Simplicity: Why Simple Linear Models Outperform Complex Machine Learning Techniques -- Case Of Breast Cancer Diagnosis," Jun. 2023, [Online]. Available: <http://arxiv.org/abs/2306.02449>
- [43] G. Zeng, "A comprehensive study of coefficient signs in weighted logistic regression," *Heliyon*, vol. 10, no. 15, Aug. 2024, doi: 10.1016/j.heliyon.2024.e35040.

- [44] S. Park, E. Ceulemans, and K. Van Deun, “Logistic regression with sparse common and distinctive covariates,” *Behav Res Methods*, vol. 55, no. 8, pp. 4143–4174, Dec. 2023, doi: 10.3758/s13428-022-02011-2.
- [45] E. Dumitrescu, S. Hué, C. Hurlin, and S. Tokpavi, “Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects,” *Eur J Oper Res*, vol. 297, no. 3, pp. 1178–1192, Mar. 2022, doi: 10.1016/j.ejor.2021.06.053.
- [46] A. F. Lukman, S. Mohammed, O. Olaluwoye, and R. A. Farghali, “Handling Multicollinearity and Outliers in Logistic Regression Using the Robust Kibria–Lukman Estimator,” *Axioms*, vol. 14, no. 1, p. 19, Dec. 2024, doi: 10.3390/axioms14010019.
- [47] P. Geurts, A. Irrthum, and L. Wehenkel, “Supervised learning with decision tree-based methods in computational and systems biology.”
- [48] O. Theobald, “Machine Learning For Absolute Beginners,” 2017.
- [49] J. R. Quinlan, “Induction of Decision Trees,” 1986.
- [50] J. Ross, Q. M. Kaufmann, and S. L. Salzberg, “Book Review: C4.5: Programs for Machine Learning,” Publishers, Inc, 1994.
- [51] M. Mahmud *et al.*, “Implementation of C5.0 Algorithm using Chi-Square Feature Selection for Early Detection of Hepatitis C Disease,” *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 6, no. 2, pp. 116–124, Apr. 2024, doi: 10.35882/jeeemi.v6i2.384.
- [52] S. Rajeswari and K. Suthendran, “C5.0: Advanced Decision Tree (ADT) classification model for agricultural data analysis on cloud,” *Comput Electron Agric*, vol. 156, pp. 530–539, Jan. 2019, doi: 10.1016/j.compag.2018.12.013.
- [53] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Chapman and Hall/CRC, 1984.
- [54] S. L. Salzberg, “Review of C4.5 Programs for Machine Learning by J. Ross Quinlan,” *Mach Learn*, vol. 16, no. 3, pp. 235–240, 1994, Accessed: Apr. 22, 2025. [Online]. Available: 10.1007/BF00993309
- [55] R. Pandya, J. Pandya, K. P. Dholakiya, and I. Amreli, “C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning,” 2015.

- [56] G. V. Kass, “An Exploratory Technique for Investigating Large Quantities of Categorical Data,” *J R Stat Soc Ser C Appl Stat*, vol. 29, no. 2, pp. 119–127, 1980.
- [57] W.-Y. Loh and Y.-S. Shih, “Split Selection Methods for Classification Trees,” *Stat Sin*, vol. 7, no. 4, pp. 815–840, 1997.
- [58] L. Breiman, “Random Forests,” *Mach Learn*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [59] Z. Sun, G. Wang, P. Li, H. Wang, M. Zhang, and X. Liang, “An improved random forest based on the classification accuracy and correlation measurement of decision trees,” *Expert Syst Appl*, vol. 237, Mar. 2024, doi: 10.1016/j.eswa.2023.121549.
- [60] H. A. Salman, “View of Random Forest Algorithm Overview,” *Babylonian Journal of Machine Learning*, vol. 2, pp. 69–79, Jun. 2024, doi: 10.58496/BJML/2024/007.
- [61] Y. Bouchlaghem, Y. Akhiat, and S. Amjad, “Feature Selection: A Review and Comparative Study,” in *E3S Web of Conferences*, EDP Sciences, May 2022. doi: 10.1051/e3sconf/202235101046.
- [62] J. A. M. Sidey-Gibbons and C. J. Sidey-Gibbons, “Machine learning in medicine: a practical introduction,” *BMC Med Res Methodol*, vol. 19, no. 1, Mar. 2019, doi: 10.1186/s12874-019-0681-4.
- [63] B. Schölkopf, A. S. Risse, A. J. Smola, K. Algorithms, B. Schölkopf, and A. J. Smola, “Support Vector Machines and Kernel Algorithms,” 2002.
- [64] B. Schölkopf, A. S. Risse, A. J. Smola, K. Algorithms, B. Schölkopf, and A. J. Smola, “Support Vector Machines and Kernel Algorithms,” 2002.
- [65] Bernhard. Schölkopf and A. J. . Smola, *Learning with kernels : support vector machines, regularization, optimization, and beyond*. MIT Press, 2002.
- [66] A. M. T. Ng, “CS229 Lecture Notes: Part V – Kernel Methods,” 2019. Accessed: Apr. 22, 2025. [Online]. Available: <https://cs229.stanford.edu/>
- [67] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, “A comprehensive survey on support vector machine classification: Applications, challenges and trends,” *Neurocomputing*, vol. 408, pp. 189–215, Sep. 2020, doi: 10.1016/j.neucom.2019.10.118.
- [68] L. Auria, R. A. M. Berlin, and R. A. Moro, “Support Vector Machines (SVM) as a Technique for Solvency Analysis,” 2008, [Online]. Available:

www.diw.de<http://www.diw.de>http://www.diw.de/english/products/publications/discussion_papers/27539.html<http://ideas.repec.org/s/diw/diwwpp.html>http://papers.ssrn.com/sol3/JELJOUR_Results.cfm?form_name=journalbrowse&journal_id=1079991

- [69] R. Akbani, S. Kwek, and N. Japkowicz, “Applying Support Vector Machines to Imbalanced Datasets,” 2004.
- [70] N. Srivastava, G. Hinton, A. Krizhevsky, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” 2014.
- [71] F. E. Ahmed, “Artificial neural networks for diagnosis and survival prediction in colon cancer,” Aug. 06, 2005. doi: 10.1186/1476-4598-4-29.
- [72] R. C. Staudemeyer and E. R. Morris, “Understanding LSTM: A Tutorial into Long Short-Term Memory Recurrent Neural Networks,” Sep. 2019. [Online]. Available: <http://arxiv.org/abs/1909.09586>
- [73] P. Siswanto and R. Rulaningtyas, “Artificial Neural Network and Its Application in Medical Disease Prediction: Review Article,” Scitepress, Jun. 2020, pp. 17–25. doi: 10.5220/0009387400170025.
- [74] S. Hochreiter, “LONG SHORT-TERM MEMORY,” *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
- [75] A. Vaswani *et al.*, “Attention Is All You Need,” Jun. 2017, [Online]. Available: <http://arxiv.org/abs/1706.03762>
- [76] L. G. Wright *et al.*, “Deep physical neural networks trained with backpropagation,” *Nature*, vol. 601, no. 7894, pp. 549–555, Jan. 2022, doi: 10.1038/s41586-021-04223-6.
- [77] N. Srivastava, G. Hinton, A. Krizhevsky, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” 2014.
- [78] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” May 27, 2015, *Nature Publishing Group*. doi: 10.1038/nature14539.
- [79] A. Shah *et al.*, “A comprehensive study on skin cancer detection using artificial neural network (ANN) and convolutional neural network (CNN),” Dec. 01, 2023, *KeAi Communications Co*. doi: 10.1016/j.ceh.2023.08.002.
- [80] S. Weerakoon and M. Assadi, “Artificial Neural Network-Driven Techno-Economic Predictions for Micro Gas Turbines (MGT) Based Energy Applications,” *Energy and AI*, vol. 20, no. 40, 2025.

- [81] M. Ibrahim *et al.*, “Artificial neural network modeling for the prediction, estimation, and treatment of diverse wastewaters: A comprehensive review and future perspective,” *Chemosphere*, vol. 362, Aug. 2024, doi: 10.1016/j.chemosphere.2024.142860.
- [82] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” Mar. 2016, doi: 10.1145/2939672.2939785.
- [83] J. Mateo, J. M. Rius-Peris, A. I. Marañña-Pérez, A. Valiente-Armero, and A. M. Torres, “Extreme gradient boosting machine learning method for predicting medical treatment in patients with acute bronchiolitis,” *Biocybern Biomed Eng*, vol. 41, no. 2, pp. 792–801, Apr. 2021, doi: 10.1016/j.bbe.2021.04.015.
- [84] K. Budholiya, S. K. Shrivastava, and V. Sharma, “An optimized XGBoost based diagnostic system for effective prediction of heart disease,” *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 7, pp. 4514–4523, Jul. 2022, doi: 10.1016/j.jksuci.2020.10.013.
- [85] K. Jane, “Ensemble Methods: Combining multiple models to improve prediction accuracy and robustness.” [Online]. Available: <https://www.researchgate.net/publication/384627805>
- [86] A. F. Jadama, A. J. Ca, and B. Jobarteh, “Ensemble Learning: Methods, Techniques, Application 4 th Modou K Toray”, doi: 10.13140/RG.2.2.28017.08802.
- [87] S. Huang, B. Wang, J. Qiu, J. Yao, G. Wang, and G. Yu, “Parallel ensemble of online sequential extreme learning machine based on MapReduce,” *Neurocomputing*, vol. 174, pp. 352–367, Jan. 2016, doi: 10.1016/j.neucom.2015.04.105.
- [88] N. F. Idris and M. A. Ismail, “A review of homogenous ensemble methods on the classification of breast cancer data,” 2024, *Wydawnictwo SIGMA-NOT*. doi: 10.15199/48.2024.01.21.
- [89] K.-W. Hsu, “Hybrid ensembles of decision trees and artificial neural networks,” in *Proceedings of the 2012 IEEE International Conference on Computational Intelligence and Cybernetics (CyberneticsCom)*, IEEE, 2012.
- [90] T.-H. Lee, A. Ullah, and R. Wang, “Bootstrap Aggregating and Random Forest,” in *Macroeconomic Forecasting in the Era of Big Data*, P. Fuleky, Ed., Springer, 2020. doi: 10.1007/978-3-030-31150-6_13.

- [91] Y. Freund and R. E. Schapire, “A Short Introduction to Boosting,” 1999. [Online]. Available: www.research.att.com/
- [92] A. K. Seewald, “Meta-Learning for Stacked Classification,” *Mach Learn*, vol. 54, no. 3, pp. 271–290, 2004.
- [93] S. M. Baik, K. S. Hong, J. M. Lee, and D. J. Park, “Integrating ensemble and machine learning models for early prediction of pneumonia mortality using laboratory tests,” *Heliyon*, vol. 10, no. 14, Jul. 2024, doi: 10.1016/j.heliyon.2024.e34525.
- [94] S. González, S. García, J. Del Ser, L. Rokach, and F. Herrera, “A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities,” *Information Fusion*, vol. 64, pp. 205–237, Dec. 2020, doi: 10.1016/j.inffus.2020.07.007.
- [95] A. A. Khan, O. Chaudhari, and R. Chandra, “A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation,” Jun. 15, 2024, *Elsevier Ltd.* doi: 10.1016/j.eswa.2023.122778.
- [96] T. Pham *et al.*, “DeepARV: ensemble deep learning to predict drug-drug interaction of clinical relevance with antiretroviral therapy,” *NPJ Syst Biol Appl*, vol. 10, no. 1, Dec. 2024, doi: 10.1038/s41540-024-00374-0.
- [97] S. Wolff, F. O’Donncha, and B. Chen, “Statistical and machine learning ensemble modelling to forecast sea surface temperature,” Sep. 2019, doi: 10.1016/j.jmarsys.2020.103347.
- [98] R. Hull *et al.*, “Simulation-based inference for parameter estimation of complex watershed simulators,” *Hydrol Earth Syst Sci*, vol. 28, no. 20, pp. 4685–4713, Oct. 2024, doi: 10.5194/hess-28-4685-2024.
- [99] Z. Wang, X. Chen, Y. Wu, L. Jiang, S. Lin, and G. Qiu, “A robust and interpretable ensemble machine learning model for predicting healthcare insurance fraud,” *Sci Rep*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-024-82062-x.
- [100] D. Berrar, “Cross-validation,” in *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, vol. 1–3, Elsevier, 2018, pp. 542–545. doi: 10.1016/B978-0-12-809633-8.20349-X.
- [101] M. Bhagat and B. Bakariya, “A Comprehensive Review of Cross-Validation Techniques in Machine Learning.”

- [102] E. Kee, J. J. Chong, Z. J. Choong, and M. Lau, “A Comparative Analysis of Cross-Validation Techniques for a Smart and Lean Pick-and-Place Solution with Deep Learning,” *Electronics (Switzerland)*, vol. 12, no. 11, Jun. 2023, doi: 10.3390/electronics12112371.
- [103] A. Zheng, *Evaluating Machine Learning Models: A Beginner’s Guide to Key Concepts and Pitfalls*. O’Reilly Media, 2015. Accessed: Apr. 22, 2025. [Online]. Available: <https://www.oreilly.com/library/view/evaluating-machine-learning/>
- [104] S. Raschka, “Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning,” Nov. 2018, [Online]. Available: <http://arxiv.org/abs/1811.12808>
- [105] V. Lumumba, D. Kiprotich, M. Mpaine, N. Makena, and M. Kavita, “Comparative Analysis of Cross-Validation Techniques: LOOCV, K-folds Cross-Validation, and Repeated K-folds Cross-Validation in Machine Learning Models,” *American Journal of Theoretical and Applied Statistics*, vol. 13, no. 5, pp. 127–137, Oct. 2024, doi: 10.11648/j.ajtas.20241305.13.
- [106] S. Prusty, S. Patnaik, and S. K. Dash, “SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer,” *Frontiers in Nanotechnology*, vol. 4, Aug. 2022, doi: 10.3389/fnano.2022.972421.
- [107] S. Szeghalmy and A. Fazekas, “A Comparative Study of the Use of Stratified Cross-Validation and Distribution-Balanced Stratified Cross-Validation in Imbalanced Learning,” *Sensors*, vol. 23, no. 4, Feb. 2023, doi: 10.3390/s23042333.
- [108] H. Ranglani, “Empirical Analysis of the Bias-Variance Tradeoff Across Machine Learning Models,” *Machine Learning and Applications: An International Journal*, vol. 11, no. 4, pp. 01–12, Dec. 2024, doi: 10.5121/mlaij.2024.11401.
- [109] G. Varoquaux, O. Colliot, and G. Varoquaux, “Evaluating machine learning models and their diagnostic value.” [Online]. Available: <https://hal.science/hal-03682454v4>
- [110] Gs. Suneetha, “Classification Evaluation Metrics in Machine learning,” JETIR, 2022. [Online]. Available: www.jetir.org/c684
- [111] J. S. Akosa, “Predictive Accuracy: A Misleading Performance Measure for Highly Imbalanced Data.”
- [112] Ž. Vujović, “Classification Model Evaluation Metrics,” *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, pp. 599–606, 2021, doi: 10.14569/IJACSA.2021.0120670.

- [113] A. Pareek, M. Huang, and P. Student, “Lecture 6: Evaluation metrics for machine learning in healthcare.”
- [114] H. M and S. M.N, “A Review on Evaluation Metrics for Data Classification Evaluations,” *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, pp. 01–11, Mar. 2015, doi: 10.5121/ijdkp.2015.5201.
- [115] M. Steurer, R. J. Hill, and N. Pfeifer, “Metrics for Evaluating the Performance of Machine Learning Based Automated Valuation Models,” 2020. [Online]. Available: www.kaggle.com/competitions
- [116] Ž. Đ. Vujović, “Classification Model Evaluation Metrics.” [Online]. Available: www.ijacsa.thesai.org
- [117] Nci and Seer, “EXTENT OF DISEASE (EOD) 2018 GENERAL CODING INSTRUCTIONS MARCH 2018.”
- [118] C. Rasola *et al.*, “Time to recurrence and its relation to survival after recurrence in patients resected for stage III colon cancer,” *Eur J Cancer*, vol. 194, Nov. 2023, doi: 10.1016/j.ejca.2023.113321.
- [119] J. H. Paik, C. G. Ryu, and D. Y. Hwang, “Risk factors of recurrence in TNM stage I colorectal cancer,” *Ann Surg Treat Res*, vol. 104, no. 5, pp. 281–287, 2023, doi: 10.4174/astr.2023.104.5.281.
- [120] S. Alinia, M. Asghari-Jafarabadi, L. Mahmoudi, S. Norouzi, M. Safari, and G. Roshanaei, “Survival prediction and prognostic factors in colorectal cancer after curative surgery: insights from cox regression and neural networks,” *Sci Rep*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-42926-0.
- [121] D. Roshani, G. Moradi, and M. A. Rasouli, “Survival Analysis of Patients with Colorectal Cancer Undergoing Combined Treatment: A Retrospective Cohort Study,” *J Res Health Sci*, vol. 23, no. 1, Mar. 2023, doi: 10.34172/jrhs.2023.107.
- [122] C. D. Blanke *et al.*, “Impact of young age on treatment efficacy and safety in advanced colorectal cancer: A pooled analysis of patients from nine first-line phase III chemotherapy trials,” *Journal of Clinical Oncology*, vol. 29, no. 20, pp. 2781–2786, Jul. 2011, doi: 10.1200/JCO.2010.33.5281.
- [123] B. Ma, Y. Li, and Q. Meng, “The predictive and prognostic value of sex in localized colorectal cancer: A SEER-based analysis,” *Transl Cancer Res*, vol. 10, no. 5, pp. 2108–2119, May 2021, doi: 10.21037/tcr-20-3421.

- [124] M. Liu *et al.*, “Marriage is a dependent risk factor for mortality of colon adenocarcinoma without a time-varying effect.” [Online]. Available: www.impactjournals.com/oncotarget
- [125] I. El Badisy *et al.*, “Risk factors affecting patients survival with colorectal cancer in Morocco: survival analysis using an interpretable machine learning approach,” *Sci Rep*, vol. 14, no. 1, 2024, doi: 10.1038/s41598-024-51304-3.
- [126] “Colorectal Cancer Early Detection, Diagnosis, and Staging.” [Online]. Available: https://seer.cancer.gov/csr/1975_2016/,
- [127] A. I. Phipps *et al.*, “Colon and rectal cancer survival by tumor location and microsatellite instability: The colon cancer family registry,” *Dis Colon Rectum*, vol. 56, no. 8, pp. 937–944, Aug. 2013, doi: 10.1097/DCR.0b013e31828f9a57.
- [128] S. Alinia, M. Asghari-Jafarabadi, L. Mahmoudi, S. Norouzi, M. Safari, and G. Roshanaei, “Survival prediction and prognostic factors in colorectal cancer after curative surgery: insights from cox regression and neural networks,” *Sci Rep*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-42926-0.
- [129] S. Hatano, H. Ishida, K. Ishibashi, K. Kumamoto, N. Haga, and I. Miura, “Identification of risk factors for recurrence in high-risk stage II colon cancer,” *Int Surg*, vol. 98, no. 2, pp. 114–121, 2013, doi: 10.9738/CC131.
- [130] April. Fritz, *International classification of diseases for oncology*. World Health Organization, 2014.
- [131] C. Ott *et al.*, “Advanced Mucinous Colorectal Cancer: Epidemiology, Prognosis and Efficacy of Chemotherapeutic Treatment,” *Digestion*, vol. 98, no. 3, pp. 143–152, 2018.
- [132] F. Maione *et al.*, “Diagnosis and management of rectal neuroendocrine tumors (Nets),” 2021, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/diagnostics11050771.
- [133] R. Sood, T. Badjatia, and P. Bhargava, “Mucinous rectal adenocarcinoma recurrence: A case report and literature review,” *Radiol Case Rep*, vol. 20, no. 4, pp. 2189–2193, Apr. 2025, doi: 10.1016/j.radcr.2025.01.046.
- [134] C.-H. Lee, S.-C. Cheng, H.-Y. Tung, S.-C. Chang, C.-Y. Ching, and S.-F. Wu, “The Risk Factors Affecting Survival in Colorectal Cancer in Taiwan,” 2018. [Online]. Available: <http://ijph.tums.ac.ir>

- [135] A. Tamm *et al.*, “Establishing a colorectal cancer research database from routinely collected health data: the process and potential from a pilot study,” *BMJ Health Care Inform*, vol. 29, no. 1, 2022, doi: 10.1136/bmjhci-2021-100535.
- [136] H. Sung *et al.*, “Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries,” *CA Cancer J Clin*, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.
- [137] Donald M. Parkin, “The evolution of the population-based cancer registry,” *Nat Rev Cancer*, vol. 6, no. 8, pp. 603–612, 2006.
- [138] Lauby-Secretan and *et al.*, “The IARC Perspective on Colorectal Cancer Screening,” *New England Journal of Medicine*, vol. 378, no. 18, pp. 1734–1740, 2018.
- [139] World Health Organization, “Saving lives, spending less: a strategic response to noncommunicable diseases,” Geneva, 2018.
- [140] F. Bray and *et al.*, “Global cancer transitions according to the Human Development Index (2008–2030): a population-based study,” *Lancet Oncol*, vol. 13, no. 8, pp. 790–801, 2012.
- [141] S. Guthmuller, V. Carrieri, and A. Wübker, “Effects of organized screening programs on breast cancer screening, incidence, and mortality in Europe,” *J Health Econ*, vol. 92, Dec. 2023, doi: 10.1016/j.jhealeco.2023.102803.
- [142] American Association for Cancer Research, “AACR Cancer Progress Report 2024,” 2024. Accessed: Dec. 15, 2024. [Online]. Available: <https://cancerprogressreport.org>
- [143] P. Nascimento de Lima, L. Matrajt, G. Coronado, A. L. Escaron, and C. M. Rutter, “Cost-Effectiveness of Noninvasive Colorectal Cancer Screening in Community Clinics,” *JAMA Netw Open*, vol. 8, no. 1, p. e2454938, Jan. 2025, doi: 10.1001/jamanetworkopen.2024.54938.
- [144] A. I. Gana *et al.*, “Cost, Cost-effectiveness and Survival Impact Assessment for A Better Management of Colorectal Cancer Patients: A Single Centre Comprehensive Analysis,” *Journal of Gastrointestinal and Liver Diseases*, Dec. 2024, doi: 10.15403/jgld-5796.
- [145] M. Nonikashvili and *et al.*, “Colorectal cancer prevention in Central, Eastern European and South Caucasus countries: a review of policy approaches,” *J Health Organ Manag*, vol. 38, no. 4, pp. 1477–7266, 2024.

- [146] M. Cloeren and et al, “Colorectal Cancer (CRC) Screening in Occupational Health Surveillance Exams Is Associated With Decreased CRC Mortality,” *Am J Ind Med*, vol. 68, no. 2, pp. 202–209, 2025.
- [147] C. Estes, M. Dehghani, A. B. Ozbay, V. Vahdat, P. J. Limburg, and D. Brooks, “Clinical, social, and economic impacts of colorectal cancer screening with the multi-target stool-DNA test: 10-year experience - a simulated study,” Aug. 08, 2024. doi: 10.1101/2024.08.07.24311643.
- [148] M. A. Preston and et al, “Health Insurance Coverage Mandates: Colorectal Cancer Screening in the Post-ACA Era,” *Cancer Prevention Research (Philadelphia)*, vol. 14, no. 1, pp. 123–130, 2021.
- [149] T. Nzeako and O. Chido-Amajuoyi, “Understanding the Impact of Numerous Cancer Prevention Guidelines on Public Adherence,” *Cancer Epidemiology, Biomarkers & Prevention*, vol. 33, no. 9, 2024.
- [150] S. Shah, P. Shridhar, and D. Gohil, “View of Control chart: A statistical process control tool in pharmacy,” *Asian J Pharm*, 2010, doi: 10.4103/0973-8398.72116.
- [151] S. M. M. Raza *et al.*, “Use of improved memory type control charts for monitoring cancer patients recovery time censored data,” *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-55731-0.
- [152] R. A. Sanusi *et al.*, “Transitions between versions of the International Classification of Diseases and chronic disease prevalence estimates from administrative health data: a population-based study,” *BMC Public Health*, vol. 22, no. 1, Dec. 2022, doi: 10.1186/s12889-022-13118-8.
- [153] N. C. Hamm, D. Jiang, R. A. Marrie, P. Irani, and L. M. Lix, “Control charts for chronic disease surveillance: testing algorithm sensitivity to changes in data coding,” *BMC Public Health*, vol. 22, no. 1, Dec. 2022, doi: 10.1186/s12889-021-12328-w.
- [154] S. Bazarbashi, H. Al Eid, and J. Minguet, “Cancer incidence in Saudi Arabia: 2012 data from the Saudi cancer registry,” *Asian Pacific Journal of Cancer Prevention*, vol. 18, no. 9, pp. 2437–2444, Sep. 2017, doi: 10.22034/APJCP.2017.18.9.2437.
- [155] N. Al Hamdan *et al.*, “Cancer registration in Saudi Arabia,” *Asian Pacific Journal of Cancer Prevention*, vol. 2, no. IACR Supplement, pp. 61–64, 2001.
- [156] S. A. Alessy *et al.*, “Cancer Research Challenges and Potential Solutions in Saudi Arabia: A Qualitative Discussion Group Study,” *JCO Glob Oncol*, no. 10, Feb. 2024, doi: 10.1200/go.23.00189.

- [157] E. Alkhalawi, "Cervical cancer in Saudi Arabia: trends and regional disparities in incidence and survival," 2021.
- [158] King Faisal Specialist Hospital & Research Centre, "Palliative Care." Accessed: Jan. 15, 2025. [Online]. Available: <https://www.kfshrc.edu.sa/en/healthcare/patients-and-visitors/palliative-care>
- [159] S. Alshammaray, B. Duraisamy, Y. Albalawi, and S. Ratnapalan, "Development of Palliative and End of Life Care: The Current Situation in Saudi Arabia," *Cureus*, Mar. 2019, doi: 10.7759/cureus.4319.
- [160] European Society for Medical Oncology, "King Faisal Specialist Hospital and Research Center – Oncology Center." Accessed: Feb. 17, 2025. [Online]. Available: <https://www.esmo.org/designated-centres/king-faisal-specialist-hospital-and-research-center-oncology-center>
- [161] S. N. Bazarbashi, "Saudi Oncology Society clinical management guidelines development," *Saudi Med J*, vol. 35, no. 12, pp. 1524–1526, 2014, Accessed: Jun. 15, 2024. [Online]. Available: <https://smj.org.sa/content/smj/35/12/1524.full.pdf>
- [162] Saudi Health Council and National Cancer Center, "Cancer Incidence Report, Saudi Arabia 2019," 2021. Accessed: May 15, 2024. [Online]. Available: <https://www.ncc.gov.sa>
- [163] O. Abulkhair, F. Al Tahan, S. Young, S. Musaad, and A. R. Jazieh, "The first national public breast cancer screening program in Saudi Arabia," *Ann Saudi Med*, vol. 30, no. 5, Sep. 2010, doi: 10.4103/0256-4947.67078.
- [164] H. O. Al-Shamsi, I. H. Abu-Gheida, F. Iqbal, and A. Al-Awadhi, *Cancer in the Arab World*. Springer Nature, 2022. doi: 10.1007/978-981-16-7945-2.
- [165] Z. A. Memish *et al.*, "Factors associated with public awareness of the Crown Health Program in the Al-Jouf Region," *J Family Community Med*, vol. 22, no. 1, pp. 31–38, Jan. 2015, doi: 10.4103/2230-8229.149586.
- [166] Z. A. Memish, A. S. Abdullah, M. Y. Saeedi, R. M. Salloum, A. J. Almadani, and O. Abid, "Methods and status of a comprehensive community-based intervention focusing on non-communicable diseases and the major risk factors in the Kingdom of Saudi Arabia. The Crown Health Project," 2013. [Online]. Available: www.smj.org.sa
- [167] M. Alattas and M. T. Alattas, "Cancer Control Priorities and Challenges in Saudi Arabia: A Preliminary projection of Cancer Burden," 2019. [Online]. Available: <https://www.researchgate.net/publication/330881945>

- [168] M. Al Asmri, M. J. Almalki, G. Fitzgerald, and M. Clark, "The public health care system and primary care services in Saudi Arabia: A system in transition," Apr. 01, 2020, *World Health Organization*. doi: 10.26719/emhj.19.049.
- [169] A. Abusanad, "Breast Cancer Stage Migration in Saudi Arabia: Examining the Influence of Screening," *Global Journal on Quality and Safety in Healthcare*, vol. 5, no. 1, pp. 24–26, Feb. 2022, doi: 10.36401/jqsh-21-15.
- [170] H. O. Al-Shamsi, I. H. Abu-Gheida, F. Iqbal, and A. Al-Awadhi, *Cancer in the Arab World*. Springer Nature, 2022. doi: 10.1007/978-981-16-7945-2.
- [171] A. Outif, "Saudi Particle Therapy Centre (SPTC) Update 2017," 2017.
- [172] G. Zacharakis *et al.*, "A 5-year evaluation of early-and late-onset sporadic colorectal cancer screening in Central Saudi Arabia," *Saudi Journal of Gastroenterology*, vol. 29, no. 2, pp. 95–101, 2023, doi: 10.4103/sjg.sjg_333_22.
- [173] N. J. Farsi, A. H. Baharoon, A. E. Jiffri, H. Z. Marzouki, M. A. Merdad, and L. A. Merdad, "Human papillomavirus knowledge and vaccine acceptability among male medical students in Saudi Arabia," *Hum Vaccin Immunother*, vol. 17, no. 7, pp. 1968–1974, 2021, doi: 10.1080/21645515.2020.1856597.
- [174] K. Salman and A. Al Saud, "Custodian of the Two Holy Mosques My primary objective is for our country to be a global model of excellence, on all fronts."
- [175] J. Jansent, "On the Statistical Analysis of Ordinal Data when Extravariation is Present," 1990. [Online]. Available: <https://academic.oup.com/jrsssc/article/39/1/75/6985415>
- [176] C. Feng, C. Feng, L. Li, and A. Sadeghpour, "A comparison of residual diagnosis tools for diagnosing regression models for count data," *BMC Med Res Methodol*, vol. 20, no. 1, Jul. 2020, doi: 10.1186/s12874-020-01055-2.
- [177] A. T. L. Lun and G. K. Smyth, "No counts, no variance: Allowing for loss of degrees of freedom when assessing biological variability from RNA-seq data," *Stat Appl Genet Mol Biol*, vol. 16, no. 2, pp. 83–93, Apr. 2017, doi: 10.1515/sagmb-2017-0010.
- [178] M. Waterhouse, I. Smith, H. Assareh, and K. Mengersen, "Implementation of multivariate control charts in a clinical setting," *International Journal for Quality in Health Care*, vol. 22, no. 5, pp. 408–414, Aug. 2010, doi: 10.1093/intqhc/mzq044.

- [179] A. Roy, R. Widjaja, M. Wang, D. Cutright, M. Gopalakrishnan, and B. B. Mittal, "Treatment plan quality control using multivariate control charts," *Med Phys*, vol. 48, pp. 2118–2126, 2021.
- [180] T. Mahmood, P. Wittenberg, I. M. Zwetsloot, H. Wang, and K. L. Tsui, "Monitoring data quality for telehealth systems in the presence of missing data," *Int J Med Inform*, vol. 126, pp. 156–163, Jun. 2019, doi: 10.1016/j.ijmedinf.2019.03.011.
- [181] A. A. Nadi, G. Celano, and S. Steiner, "Investigating the performance of the Phase II Hotelling T2 chart when monitoring multivariate time series observations," Jan. 2025, [Online]. Available: <http://arxiv.org/abs/2501.11649>
- [182] M. ; N. S. T. A. Khedmati, "Phase II monitoring of general linear profiles in the presence of between-profile autocorrelation," *Qual Reliab Eng Int*, vol. 32, pp. 443–452, 2016.
- [183] R. E. McAuliffe, "Durbin–Watson Statistic," in *Wiley Encyclopedia of Management*, C. L. ; M. R. E. Cooper, Ed., Wiley, 2015.
- [184] R. Noorossana and A. ; S. P. Amiri, "On the monitoring of autocorrelated linear profiles," *Commun Stat Theory Methods*, vol. 37, pp. 425–442, 2008.
- [185] S. D. Grimshaw, "Constructing control charts for autocorrelated data using an exhaustive systematic samples pooled variance estimator," 2023.
- [186] X. Ma, X. Ma, L. Zhang, J. Hu, and A. Palazoglu, "A model-free approach to reduce the effect of autocorrelation on statistical process control charts," *J Chemom*, vol. 32, no. 12, 2018.
- [187] R. A. Sanusi *et al.*, "Transitions between versions of the International Classification of Diseases and chronic disease prevalence estimates from administrative health data: a population-based study," *BMC Public Health*, vol. 22, no. 1, Dec. 2022, doi: 10.1186/s12889-022-13118-8.
- [188] W. Hoeffding, "A Class of Statistics with Asymptotically Normal Distribution," vol. 19, no. 3, pp. 293–325, 1948.
- [189] M. Khedmati and S. T. A. Niaki, "Phase II monitoring of general linear profiles in the presence of between-profile autocorrelation," *Qual Reliab Eng Int*, vol. 32, no. 2, pp. 443–452, Mar. 2016, doi: 10.1002/qre.1762.
- [190] D. J. Hauck, G. C. Runger, and D. C. Montgomery, "Multivariate statistical process monitoring and diagnosis with grouped regression adjusted variables," *Commun Stat Simul Comput*, vol. 28, pp. 309–328, 1999.

- [191] M. Erfanian, B. Sadeghpour Gildeh, and A. M. Reza, “A new approach for monitoring healthcare performance using generalized additive profiles,” *J Stat Comput Simul*, vol. 91, pp. 167–179, 2021.
- [192] D. C. Montgomery, *Introduction to Statistical Quality Control*, no. 3. Wiley, 2009. Accessed: Apr. 22, 2025. [Online]. Available: <https://www.wiley.com/en-us/Introduction+to+Statistical+Quality+Control%2C+6th+Edition-p-9780470169926>
- [193] A. Mazidimoradi, F. Hadavandsiri, Z. Momenimovahed, and H. Salehiniya, “Impact of the COVID-19 Pandemic on Colorectal Cancer Diagnosis and Treatment: a Systematic Review,” Mar. 01, 2023, *Springer*. doi: 10.1007/s12029-021-00752-5.
- [194] C. Maringe *et al.*, “The impact of the COVID-19 pandemic on cancer deaths due to delays in diagnosis in England, UK: a national, population-based, modelling study,” *Lancet Oncol*, vol. 21, no. 8, pp. 1023–1034, Aug. 2020, doi: 10.1016/S1470-2045(20)30388-0.
- [195] L. Frazzoni, E. Dajti, and B. Silviu Ungureanu, “The impact of the COVID-19 pandemic on colorectal and gastric cancer diagnosis, disease stage and mortality.” [Online]. Available: www.cbs.gov.il
- [196] W. H. Woodall, “The Use of Control Charts in Health-Care and Public-Health Surveillance,” *Journal of Quality Technology*, vol. 38, pp. 89–104, 2006, Accessed: May 15, 2024. [Online]. Available: <https://www.asq.org/pub/jqt>

Vitae

Name: Maha Abdel hakeem Abdel Hameed said

Nationality: Egyptian

Date Of Birth: September 27, 1986

Email: maha_idsc@yahoo.com

Phone: (+966) 533964548

Address: Building No. 7617, King Abdulaziz Road, Al Khobar Al Shamalia District, Al Khobar 34425, Saudi Arabia

Education:

Master of Science: Applied Statistics

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

September 2023 - May 2025

Bachelor of Science: Statistics

Faculty of Economics and Political Science, Cairo University, Egypt

September 2004 - June 2007

Professional Experience:

Statistical Researcher

Information and Decision Support Center (IDSC), Cabinet of Egypt, Cairo

June 2008 – April 2016 (8 years)

- I. Developed statistical models and forecasts for 27 national-level projects, spanning economic, social, and labor market domains.
- II. Designed and analyzed large-scale surveys and databases to support government policy planning and evaluation.

- III. Performed advanced data preparation, cleaning, and transformation for high-impact research outputs.
- IV. Created customized data visualization tools and dashboards to communicate insights to decision-makers.
- V. Improved forecasting accuracy using time series models that accounted for seasonality and trends.
- VI. Authored technical reports and briefs tailored for non-expert audiences, including ministers and senior advisors.
- VII. Managed national statistical datasets and ensured proper documentation for long-term use and collaboration.

Research Contributions and Publications:

As part of my role at IDSC, I contributed to several research reports and policy briefs in both Arabic and English. Selected contributions include:

- Labor market forecasting methods study, October 2008.
- Labor market performance in Egypt study, October 2008.
- The conditions of Egyptian workers in Jordanian labor market and the job opportunities available study, October 2008.
- Training centers in Egypt: Towards a new methodology for data collection study, July 2008.
- Forecasting Labor Demand in Egypt (2006/2007-2011/2012) (Arabic and English version) study, March 2009.
- Towards building an integrated information system for the labour market in Egypt study, Oct 2009.
- The analysis of labor market, wages and workers indicators in Egypt, "a practical study on the communications and information technology sector study, Oct 2009.
- The implication of the global financial crisis of tourism in Egypt (Arabic and English version) study, 2010.
- building an information system for the labour market by using Dev- Info Program, 2010.
- Employment-Intensive Investments study, November 2011.
- International experience in the field of freedom of information laws and the possibility of application in Egypt report, June 2011.
- Monitoring the successful experiences of the employer's participation in the enhancement of technical education in Egypt report, December 2011.
- Towards enhancing the freedom of information in Egypt report, January 2012.
- The Role of Trade Union in Labor Dispute report, June 2012.

- the private sector and academics vision for graduates of the higher education system in Egypt report, September 2012.
- Trends of the Egyptian labor market after the revolution of 25 January 2011 report, Feb 2014.
- The role of civil society in the rehabilitation of youth to the labor market report, June 2014.
- Employment and working conditions in the information and communication technology sector, Informative Report, April 2010.
- Statistics and information ... The right of the citizen the basis for furthermore development, Informative Report, October 2010.
- Freedom of information laws... International experiences and the current situation in Egypt, Informative Reports, June 2011.
- The reality of education in Egypt, Informative Reports, March 2013.
- Labour market performance in Egypt (1998-2006), Policy Briefs, may 2009.
- The analysis of labor market, wages and workers indicators in Egypt, "a practical study on the communications and information technology sector, Policy Briefs, April 2010.
- The implication of the global financial crisis of tourism in Egypt, Policy Briefs, may 2010.
- Change in the pattern of the structure of occupations in the Egyptian labor market between 1996 and 2006, Policy Briefs, August 2010.
- Employment-Intensive Investments, Policy Briefs, may 2014.

Seminar and Conference Participation:

- I. I have actively participated in and helped organize a number of national conferences and consultative workshops focusing on labor economics, technical and vocational education, employment policy, and development strategies. Notable examples include:
- II. The Third Conference of the Egyptian Observatory for Education, Training, and Employment.
- III. Workshops on Employment-Intensive Investment Strategies (EII), Decent Work in Egypt, and Freedom of Information Policies in the MENA Region.

Training & Professional Development

- **Statistical Research Training**, Youth Association People of Development, Egypt — 2006

Participated in a 2-month training program focusing on basic data collection and statistical reporting.

- **Statistical Research Internship**, Poll Center, Information and Decision Support Center (IDSC), Egypt — *Jul–Sep 2007*
Supported survey design, data entry, and analysis for national opinion polling projects.
- **Statistical Research Internship**, Egyptian Observatory for Education, Training, and Employment, IDSC — *Jan–Jun 2008*
Conducted labor market data analysis and contributed to employment-related reports.
- **Workshop on Impact Evaluation**, Perspectives on Impact Evaluation, Egypt — *Mar 28 – Apr 2, 2009*
Attended a technical workshop covering frameworks and tools for evaluating social and economic programs.
- **Poverty Targeting and STATA Application Course**, IDSC — *2010*
Gained hands-on experience using STATA for data analysis and targeting mechanisms for social support programs.
- **Capacity Building Program**, IDSC in collaboration with InWEnt (Germany) — *2011*
Participated in advanced training on public policy formulation, implementation, and evaluation.
- **Public Policy Support Training**, IDSC — *2012*
Focused on policy impact assessment and evidence-based decision-making.
- **Training on Poverty Indicators and Targeting Methods**, IDSC — *Feb 8–20, 2014*
Covered multidimensional poverty measurement, targeting strategies, and application in national social protection programs.