Instituto Tecnologico y de Estudios Superiores de Monterrey

Monterrey Campus

School of Engineering and Sciences

Risk of Breast Cancer in the Mexican Population: A Radiomics Approach

A thesis presented by

Yareth Lafarga Osuna

Submitted to the
School of Engineering and Sciences
in partial fulfillment of the requirements for the degree of

Master of Science

in

Computer Science

Monterrey, Nuevo León, May, 2023

Instituto Tecnologico y de Estudios Superiores de Monterrey

Campus Monterrey

The committee members, hereby, certify that have read the thesis presented by Yareth Lafarga Osuna and that it is fully adequate in scope and quality as a partial requirement for the degree of Master of Science in Computer Sciences.

Dr. José Gerardo Tamez Peña
Tecnológico de Monterrey
Principal Advisor

Dr. Alejandro Santos Díaz
Tecnológico de Monterrey
Committee Member

Dr. Juan Emmanuel Martínez Ledezma
Tecnológico de Monterrey
Committee Member

Dr. Benjamin Castañeda
Pontificia Universidad Católica del Perú
Committee Member

Dr. Ruben Morales Menéndez
Associate Dean of Graduate Studies
School of Engineering and Sciences

Monterrey, Nuevo León, May, 2023

Declaration of Authorship

I, Yareth Lafarga Osuna, declare that this thesis titled, Risk of Breast Cancer in the Mexican Population: A Radiomics Approach and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Yareth Lafarga Osuna
Monterrey, Nuevo León, May, 2023

Dedication

With profound gratitude and unwavering devotion, I wholeheartedly dedicate this thesis to my beloved family, whose steadfast support has been the bedrock of my academic journey.

First and foremost, I extend my most profound appreciation to my extraordinary mother, whose constant encouragement and unwavering love have illuminated my path at every step. Despite the divergence between my chosen paths and your ideals, you have remained my ultimate source of motivation in pursuing and completing my master's degree. Your patience, unwavering confidence, and unparalleled intelligence have been an endless source of inspiration.

To my dear father, I express my eternal gratitude for your unwavering belief in me and your steadfast support that has nurtured my wildest ideas and fueled my adventurous spirit. The solid foundation you laid through my early education ignited a fire within me, propelling me toward pursuing advanced research. I am forever indebted to both of my parents for equipping me with the tools and driving my thirst for knowledge.

To my treasured brother, my constant companion throughout the formative years of our childhood, I extend my heartfelt thanks. You have been my guiding light, unwavering supporter, and loyal playmate. Your consistent encouragement, assistance in my daring pursuits, and shared laughter have shaped my resilience and work ethic. Your role as my protector and mentor holds a special place in my heart.

To my inspiring sister, a beacon of strength and determination, I am deeply grateful for your unwavering support. Your example has ignited a passion for continued learning within me. Thank you for your persistent presence, uplifting words of encouragement, and genuine concern. Your steadfast support has provided me with immeasurable comfort throughout this journey.

A particular portion of this dedication is devoted to all breast cancer patients and survivors, with heartfelt recognition to Lupita. Your inspiring testimony has served as a wellspring of motivation, propelling me to develop this research.

Lastly, I dedicate this work to little Yareth, who dreamed of becoming a scientist. Together, we have taken a step forward and conquered all obstacles that stood in our way, most notably the hurdles within my mind.

I express profound gratitude to my family, friends, survivors, and partner for your unwavering presence, belief, and endless inspiration. This thesis is a testament to the collective love and support that propelled me forward on this transformative academic odyssey.

Acknowledgements

I want to sincerely appreciate the individuals and institutions who have played a crucial role in making this academic journey possible. First and foremost, my heartfelt thanks go to the scholarship providers and the esteemed institution, Tecnologico de Monterrey, for their unwavering support.

Furthermore, I thank CONAHCyT for generously providing the tuition that allowed me to live and study in Monterrey. Their contribution has enabled me to focus on my research and academic pursuits.

I would also like to acknowledge the invaluable support and assistance from San José's Hospital for granting me access to the dataset used in this research. Their collaboration has been integral to the successful execution of this work.

My most profound appreciation goes to my advisor, Dr. Jose Tamez, for believing in my abilities and entrusting me with this research endeavor. His guidance, unwavering faith, motivation, extensive knowledge, and support have been invaluable throughout this journey. I am also grateful to the committee members for their valuable feedback and suggestions that have helped enhance the quality of my work.

I owe a debt of gratitude to my friends, who have been a constant source of motivation, pushing me to excel and providing unwavering support. In particular, I want to express my special thanks to my friend Eduardo, whose persistent presence and encouragement have carried me through the ups and downs of this academic pursuit. Additionally, I am grateful to my roommate, Diana, who has always been available to lend an ear, even when tired, and has been a steadfast companion in listening to my thoughts and adventures. My partner, I am deeply grateful for your unwavering support, patience, moments of fun, and boundless love.

I extend my heartfelt gratitude to my beloved pets, Keki, Chikis, and especially Kobe, for their unconditional love and unwavering emotional support throughout this journey. Their presence has been a constant source of comfort and solace.

To all who have played a part in this journey, whether through support, guidance, or companionship, I offer my heartfelt appreciation. Your contributions have been invaluable, and this thesis stands as a testament to our collective efforts and unwavering dedication.

Risk of Breast Cancer in the Mexican Population: A Radiomics Approach

by

Yareth Lafarga Osuna

Abstract

Breast cancer is a significant global health concern, especially among women, with rising incidence rates in specific populations. Low screening rates contribute to this alarming trend, emphasizing the need to improve breast cancer risk prediction and enhance screening outcomes. This thesis explores the potential of image-based models and machine learning techniques to address limitations in traditional risk assessment models and leverage the rich information available in mammography images. A larger data set, including diverse cases with breast cancer diagnoses, is recommended to improve accuracy and unreliability. In addition, extracting additional image-based features to characterize breast anatomy could provide valuable insights. The outcomes of this research can contribute to personalized medicine approaches and improve breast cancer risk prediction, leading to early detection, timely interventions, and improved patient outcomes. This study showed successful segmentation and extraction of 78 features per image (first and second order), and the methodology's performance with a machine learning Cox model achieved an AUC of 0.76. Furthermore, the Kaplan-Meier curve significantly differed between the low-risk and high-risk groups. The advantage of using a Cox model is its ability to identify the most discriminative features, which in this case were three features associated with the physiological characteristics of the patients. This thesis provides a roadmap for further investigation, emphasizing the importance of larger data sets, technique refinement, and exploration of population-specific characteristics to develop more effective breast cancer screening and prevention strategies.

Keywords: Radiomics, Breast Cancer (BC), Prognosis, Machine Learning.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Breast cancer (BC) is the most prevalent cancer among women. It has the highest incidence worldwide [148], with an estimated 2.3 million new cases each year, accounting for 11.7% of all cancer cases and 685,000 fatalities, accounting for 6.9% of all cancer deaths [132]. Currently, invasive Breast Cancer represents the first cancerous death among Mexican and Hispanic women [95]. Despite having a lower overall incidence and mortality of breast cancer, their incidence rates are rising quicker [149], and their mortality drops are slower than those of other populations. As this population ages and gets more acculturated, it is anticipated that the burden of breast cancer will increase[95]. Low screening mammography rates among Hispanic women contribute to these trends in incidence and mortality [126]. Therefore, it's crucial to describe their screening results more fully.

Each country has the norm to advise women from a certain age to start the screening for BRCA [31]. In Mexico, the NOM-041-SSA2-201 suggests women from 40 to 69 years have regular screening with mammography every two years [33]. Mammography is a reliable screening method using low-energy X-rays to identify and treat breast cancer in its earliest stages. Digital mammography is a type of breast imaging that collects and analyzes mammographic pictures using digital technology [45]. Improved image quality, quicker results, less radiation exposure, digital storage and retrieval, tomosynthesis capabilities, improved image editing, and telemedicine capabilities are some benefits of digital mammography [102]. These advantages make digital mammography a useful tool for breast imaging and help with the early identification and detection of breast cancer.

Mammography interpretation is typically performed by a radiologist who examines mammographic images for abnormalities, such as masses or calcifications, that may indicate breast cancer or other conditions and evaluates their propensity to be benign or malignant [105]. BIRADS (Breast Imaging Reporting and Data System) is a standard medical system to classify and submit findings from mammography and other breast imaging studies to standardize and communicate breast imaging findings [12]. It employs a scale from 0 to 6, with 0 indicating an evaluation that couldn't be completed, 5 indicating a highly suspicious abnormality, and 6 representing a confirmed breast cancer diagnosis. BIRADS helps provide consistent and standardized reporting, guiding further evaluation and management decisions for patients with breast abnormalities, and is an essential tool for early breast cancer detection and diagnosis. The radiologist may also recommend additional imaging studies, such as mammographic views or breast ultrasounds, if an additional evaluation is required [105]. The radiologist documents

the final interpretation of the mammogram, including a description of any aberrant findings and their characteristics, in a written report. The interpretation of mammograms requires specialized training and expertise, and the accuracy depends on several factors.

Forecasting the probability of developing breast cancer is vital for avoiding high-risk people and achieving early detection. It has been demonstrated that an early diagnosis increases the likelihood of survival [30]. An early BC discovery implies that the treatment must begin promptly. A good quantity of BC risk assessment methods was developed, and one of the most popular is the Gail Model [149]. The Gail model is a Breast Cancer Risk Assessment Tool (BCRAT), one of the most examined models according to [146]. It evaluates a prospective absolute risk calculation and the modification of the associated hazard. In this model, hormonal and pathologic data are the most significant predictors of risk. Another breast cancer risk assessment method is the International Breast Cancer Intervention Study (IBIS) model, sometimes called the Tyrer-Cuzick model, which is a more thorough model considering additional characteristics besides those in the Gail model [80]. Breast density, hormonal and reproductive history, and genetic abnormalities (such as BRCA1/BRCA2) are all considered. It has been demonstrated that these models are an appropriate instrument for patients who follow the recommended mammography screening protocol [27].

The current validated models for BC risk prediction vary in complexity, validation, and applicability [71]. Data availability, assessment purpose, and the population being assessed may influence the selection of the breast cancer risk assessment model [19]. Incomplete risk prediction, uncertainty and variability in estimates, limited generalizability, lack of incorporation of new risk factors, limited usefulness for certain populations, ethical and psychological considerations, and clinical decision-making complexity are some of the limitations and potential drawbacks of breast cancer risk assessment models like the Gail and Tyrer-Cuzick. The limitations of these models should be considered when interpreting and applying their conclusions because they are tools that provide estimates of breast cancer risk based on existing data and assumptions [71]. There have been several improvements and validation to increase the accuracy of the performance of the traditional risk assessment models, like the most used in medical practice: The Gail Model [27]. The enhancements to develop higher-precision models are focused on adapting a risk assessment tool that combines hormonal and pathological information.

Image-based models have emerged, exposing a different perspective for breast cancer evaluation [105]. The increasing use of techniques from machine learning to interpret mammography pictures has opened up new doors for improving breast cancer risk prediction[53]. Most of the information that goes into traditional risk prediction models for breast cancer comes from the patient, including their age, family medical history, and hormonal factors [19]. Although these models have helped identify those at a higher risk, there is room for improvement in their accuracy. Emerging models for determining breast cancer risk are embracing new methods, such as image-based models that use mammography data [96]. One example of this is the trend. These image-based methods extract new features from mammography pictures, supplying risk assessors with more comprehensive information. These novel models can improve early diagnosis and individualized intervention tactics, resulting in more successful breast cancer prevention and therapy.

A machine-learning image-based model typically involves several steps: feature extraction, representation, model training, evaluation, and deployment [70]. Feature extraction is a

crucial step that involves identifying and extracting relevant visual patterns from the raw image data. The quality of these features dramatically impacts the model's ability to learn and make accurate predictions. Good input features are essential as they ensure accurate representation of visual patterns, reduce data dimensionality, facilitate generalization to unseen data, provide interpretability, and allow adaptability to different problem domains [147]. Radiomics is a field that involves extracting quantitative features from medical images 100, such as X-ray images, CT scans, or MRI scans, to characterize tumor phenotypes. Radiomics features can capture complex tumor characteristics that are not visually apparent and, therefore, can provide valuable information for machine learning models [90]. Incorporating radiomics features in a machine learning model can enhance its predictive performance and aid in diagnosing, treating, planning, and outcome prediction in various medical applications, such as cancer prognosis, response to therapy, and personalized medicine [74].

1.1 Problem Definition and Motivation

Breast cancer is currently a well-recognized and intensively investigated type of cancer. Although breast cancer has most likely been there for a long time, our knowledge of it as a medical problem has changed over the years. Breast cancer was first described in writing in ancient Egypt circa 1600 BCE [40]. Nevertheless, our understanding of breast cancer has evolved throughout the past few centuries, including our knowledge of its causes, risk factors, diagnosis, and treatments. In the 19th century, the French physician Jean Louis Petit was the first to describe breast cancer as a local disease that could spread to other body parts. In the 20th century, advancements in medical technology, such as X-rays and mammography, improved the early detection of breast cancer [88]. In the mid-20th century, the development of chemotherapy and hormonal therapy revolutionized breast cancer treatment.

Breast cancer is the most common female malignancy, with 2.3 million diagnoses in 2020 [87]. Variables affect breast cancer statistics and rates. Breast cancer prevalence varies by nation, gender, age, and access to healthcare [153]. Early detection, screenings, and risk factor awareness can lessen the impact of breast cancer on individuals and communities [87]. Developments have greatly improved the early diagnosis and treatment of breast cancer in medical technology, including X-rays, mammography, chemotherapy, and hormone therapy. Breast cancer's diagnostic, treatment choices, and outcomes are constantly improving [88].

For prevention, individualized BC risk assessment models were created. Predicting BC risk helps patients and doctors make informed decisions [71]. A high-risk patient may have more frequent checkups to prevent early diagnosis than the national recommendation. Low-risk women will have longer checkups. Thus, BC risk assessment leads to customized screening and improves patient survival. Breast Cancer Risk Assessment models emerged 30 years ago. Using statistical methods, traditional approaches evaluate age, family history, and other attributes. The first model developed was the Gail et al. [44], several modifications were applied to increase its performance, and it's the most-cited among traditional models. Other traditional risk assessment models are the Rosher and Colditz model [156], the Breast Cancer Surveillance Consortium (BCSC) [139], IBIS [140] and Eriksson et al. [36] among others.

Traditional approaches rely on patient answers [19]. Thus, risk variables like family history may not be appropriately handled. Additionally, because of low accuracy, the diverse

population may impact the applicability of the model, and data input depends on the patient; it is not feasible to select a standard technique from the traditional models that accurately predict individual risk of developing Breast Cancer [86]. Current breast cancer risk assessment tools' limitations characterize this research's problem statement. These limitations highlight the need for improved risk assessment approaches, considering more comprehensive and accurate data to better identify individuals at risk for Breast Cancer. These tools heavily rely on personal information[27], but the attributes evaluated by the models need to be improved, resulting in low sensitivity and discriminative accuracy.

Artificial Intelligence (AI) methods like Machine Learning (ML) Algorithms are being used more to find patterns and mathematical linkages in complex data sets [78]. Machine Learning has improved cancer prediction accuracy by 15-20% recently [101]. ROC AUC ("Area Under the Curve" of the "Receiver Operating Characteristic" curve) is one metric to measure ML algorithm performance [41]. Feature extraction, dimensional reduction, and data feature or attribute selection boost algorithm performance [70]. Database type determines the best ML approach.

As mentioned, feature extraction and selection are needed to improve the performance of a machine learning algorithm. This issue can be addressed with Radiomics. Radiomics is defined as the translation of medical images to structured numerical data [141]. It uses a range of attributes, such as geometry, strength, and texture, determined from radiological images to capture different imaging patterns. These techniques can be implemented in multiple variations for prediction, prognosis, monitoring, and treatment response assessment[2]. These techniques can be divided into two main groups: features-based and Deep Learning-based [1]. The features-based extract is a set of numerical features from a segmented region. The advantages of this group are that they do not need large data sets, and they can be implemented in short computation time [90]. On the other hand, Deep Learning-based radiomics techniques typically use Convolutional Neural Networks to find the most critical characteristics from radiological images. The disadvantage of these techniques is the interpretability, the need for larger data sets, and a more extended period of computational time used in their implementation.

There have been works combining machine learning and radiomics to detect the presence of different diseases or other medical applications. However, the innovation of this research is the implementation of these techniques to develop a Breast Cancer Risk Assessment model. The use of Machine Learning to classify and features-based Radiomics for digital mammography attribute extraction highlights the advantages of both techniques, the applicability of seeking a correlation in image-based features almost unpredictable by human perception with a higher accuracy (ROC AUC) performance.

1.2 Hypothesis and Objectives

1.2.1 Hypothesis

Current Breast Cancer risk assessment models can improve performance by combining radiomic information and machine learning techniques from negative screening digital mammography images.

1.2.2 Objectives

General objective:
Develop an image-based model that can predict a person's risk of acquiring breast cancer by applying radiomics features and machine learning techniques.

The specific objectives to achieve this research work are:

- Review and apply pre-processing techniques to ensure consistency across the dataset's images.
- Utilize segmentation methods to accurately identify and isolate the most significant regions in the mammogram.
- Investigate and select the most relevant and informative radiomics features that have been shown to correlate with breast cancer risk, including texture, shape, and intensity features.
- Develop an image-based breast cancer risk assessment model that utilizes radiomics features and machine learning for prediction.
- Evaluate the algorithm's performance that uses radiomics features and machine learning to predict breast cancer risk within the Mexican population.

1.3 Solution Overview

The development of this project is in collaboration with "El Centro de Cáncer de mama del Hospital San José". The database contains patient information and mammograms from around 25,000 patients. The database includes annotations made by the department of a specialized radiologist. The characteristics of this data set and the collaboration with "El Centro de Cáncer de mama del Hospital San José" made the project a reachable and helpful tool for the current screening in Mexico.

The workflow of the solution overview is presented in figure 1.1. The initial step in the workflow involved inputting data from the San Jose hospital into the solution model. A filtering step was then executed to review the cases processed to ensure the necessary information was present. Specifically, four mammography views were identified as essential for processing: right craniocaudal (CC), right mediolateral oblique (MLO), left CC, and left MLO. Subsequently, the data underwent a preprocessing stage involving several steps. The images were normalized, rescaled, and mirrored to account for the fact that each screening case contained four mammography views. Next, each image was segmented into three regions: control, dense, and non-dense, which facilitated the extraction of features from the raw image. Wavelet decomposition of the original image was then carried out for three different levels, followed by additional feature extraction from the magnitude of each level wavelet decomposition. Overall, this approach aimed to identify and extract features from mammography images to predict breast cancer risk.

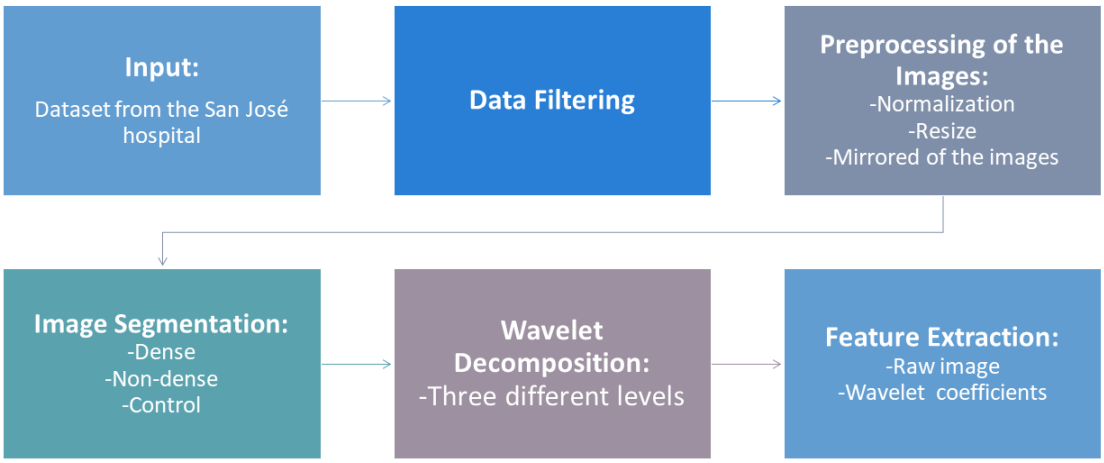

Figure 1.1: Solution Overview workflow diagram

1.4 Scientific Contributions

- The study contributes by extracting radiomics features from screening mammographies. Radiomic feature extraction highlights the importance of leveraging advanced imaging analysis techniques to derive meaningful information for breast cancer risk prediction that can correlate to biological and physical characteristics.
- The research contributes by utilizing machine learning Cox models focused on survival analysis. The implementation of machine learning Cox models focused on survival analysis demonstrates the effectiveness of these models in identifying the most discriminant features for breast cancer risk prediction.
- Implementing the proposed models is a crucial starting point for further research in the field. This work represents one of the initial investigations into the practical application of radiomic features in predicting the risk of breast cancer. Implementing these models establishes a solid foundation for future studies and advancements in personalized medicine approaches for managing breast cancer.
- The research stresses the importance of increased dissemination of similar studies to foster a breast cancer screening culture and encourage greater participation in dataset collection; this contributes to advancing research and clinical practice in breast cancer management. how which radiomic features can correlate with breast cancer risk.

1.5 Thesis Outline

Breast cancer is a prevalent disease affecting women worldwide, including in Mexico. To improve early detection and identify risk factors, this thesis explores the use of radiomics in assessing the risk of breast cancer in the Mexican population. The thesis is structured into six chapters, including an introduction that outlines the importance of the study and its objectives. The literature review examines the current knowledge regarding breast cancer risk factors, radiomics, and related diagnostic tools. The methodology chapter outlines the steps developed for this study. The results chapter presents the study's findings, including identifying potential risk factors and the effectiveness of the proposed model in predicting breast cancer risk. The discussion chapter interprets the results and explores the study's implications for breast cancer risk assessment. Finally, the conclusion chapter summarizes the study's key findings and offers suggestions for future research. Through this thesis, we aim to contribute to understanding breast cancer risk factors and developing more effective risk assessment tools for the Mexican population.

Chapter 2

Literature Review

According to the World Health Organization, in 2020, there were 2.3 million women diagnosed with breast cancer and 685,000 deaths globally [106]. In the United States, breast cancer accounts for about 30% of all new cancer cases in women each year[18]. The American Cancer Society estimates that in 2022, about 287,850 new cases of invasive breast cancer are expected to be diagnosed in women in the U.S., along with 51,400 new cases of non-invasive (in situ) breast cancer [18].

Since 2006, breast cancer has been the leading cause of cancer mortality in Mexican women, accounting for 14% of cancer-related deaths [25]. It is estimated that in 2018, breast cancer accounted for 29.5% of all new cancer cases in women in the country. Additionally, approximately 27,000 new breast cancer cases are diagnosed each year in Mexico. In 2021, approximately 7.9 thousand women died from breast cancer in Mexico [93]. GLOBOCAN's prediction that by 2030, 24,386 women will be diagnosed and 9,778 (40%) will die with breast cancer in Mexico makes this disease a substantial challenge for the health-care system [25]. Breast cancer incidence in Mexico has increased over the past few decades, partly due to lifestyle and reproductive patterns changes [116]. The incidence rates are higher in urban areas than in rural areas, and the disease tends to affect women younger than in other countries [25]. Mortality rates from breast cancer in Mexico have decreased in recent years, likely due to early detection and treatment improvements. However, breast cancer remains a significant public health issue in Mexico. More efforts are needed to improve access to breast cancer screening, diagnosis, and treatment services, especially for underserved populations.

2.1 Breast Cancer

Breast cancer begins in the breast cells, where abnormal cells multiply uncontrollably and create a lump [29]. It is categorized into different types, including ductal and lobular carcinoma, and the degree of its spread is determined by staging [117]. Staging allows healthcare professionals to determine the most effective treatment approach. It ranges from stage 0 to stage IV, with 0 indicating non-invasive breast cancer and IV indicating advanced cancer that has spread to other areas of the body [6].

- **Stage 0**: This refers to the non-invasive phase of a tumor, signifying that both malignant and benign cells remain confined within the specific area of the breast where the tumor

originates. There is no indication of these cells invading the adjacent tissues. An example of this stage of tumor development is ductal carcinoma in situ (DCIS) [13].

- **Stage 1**: This stage is defined as invasive breast carcinoma, with the potential for microscopic invasion. It is divided into two subcategories: stage 1A and stage 1B. Stage 1A represents a tumor measuring up to 2 cm without any lymph node involvement, while stage 1B is characterized by small groups of cancer cells larger than 0.2 mm in a lymph node [122].

- **Stage 2**: This stage is divided into two subcategories: 2A and 2B. Stage 2A is characterized by a tumor in axillary lymph nodes or sentinel lymph nodes, with no tumor detected in the breast. The tumor can range from smaller to larger than 2 cm, but not exceeding 5 cm. On the other hand, stage 2B indicates that the tumor may be larger than 5 cm, yet it does not extend to the axillary lymph nodes [98].

- **Stage 3**: It is categorized into three subcategories: 3A, 3B, and 3C. Stage 3A is characterized by the absence of a tumor in the breast but the presence of one in 4-9 axillary lymph nodes or sentinel lymph nodes. Stage 3B refers to a tumor of any size that has caused swelling or an ulcer on the breast skin and has spread to up to 9 axillary lymph nodes or sentinel lymph nodes. Stage 3B may be classified as inflammatory breast cancer, which typically presents with red, warm, and swollen breast skin. Stage 3C, on the other hand, involves the spread of the tumor to 10 or more axillary lymph nodes, including those above and below the clavicle [66].

- **Stage 4**: This stage represents the advanced and metastatic phase of cancer, characterized by the spread of cancer to other organs within the body, such as the lungs, bones, liver, brain, and more[104].

2.1.1 Risk of Breast Cancer

Breast cancer is generated by changes, or mutations, in the DNA of breast cells. These mutations can occur spontaneously or may be caused by external factors such as exposure to radiation, certain chemicals, or hormones. These mutations affect the genes that regulate cell growth, division, and death. When the genes that promote cell growth and division are overactive, or the genes that inhibit cell growth and division are inactive, the cells can grow and divide abnormally and form a cancerous mass, also known as a tumor [7]. If the cancer cells are not detected and treated, they may continue to grow and spread to nearby tissues or even to other parts of the body through the bloodstream or lymphatic system.

Breast cancer is a complex disease with unknown causes, but researchers have identified several risk factors. These include increasing age, reproductive factors like early menarche and late menopause, mammographic density, genetic factors, and family history (observed in about 9% of cases) [77]. Although potential risk factors like dietary fat intake, exposure to solvents and pesticides have been suggested, the evidence for them remains inconclusive. Various studies have explored the factors that women with breast cancer attribute to their condition, including psychosocial factors like stress, physical trauma to the breast, chemical exposure,

proximity to electronic equipment or overhead power lines, viral or bacterial infection, and even bad luck. However, there is limited evidence to support these attributions [77].

The work of Lukasiewicz et al. discusses the various risk factors associated with breast cancer. These risk factors can be divided into two categories: modifiable and non-modifiable [87]. Lifestyle choices such as diet and exercise can change modifiable risk factors. Non-modifiable risk factors like age and genetics cannot be changed. Table 2.1 details the criteria stated by the authors regarding the non-modifiable and modifiable risk factors. Understanding these risk factors can help individuals make informed decisions about their health and take steps to reduce their risk of developing breast cancer.

Non-Modifiable Factors	Modifiable Factors
Female sex	Hormonal replacement therapy
Older age	Diethylstilbestrol
Family history (of breast or ovarian cancer)	Physical activity
Genetic mutations	Overweight/obesity
Race/ethnicity	Alcohol intake
Pregnancy and breastfeeding	Smoking
Menstrual period and menopause	Insufficient vitamin supplementation
Density of breast tissue	Excessive exposure to artificial light
Previous history of breast cancer	Intake of processed food
Non-cancerous breast diseases	Exposure to chemicals
Previous radiation therapy	Other drugs

Table 2.1: **Modifiable and non-modifiable breast cancer risk factors.** Note. Adapted from Breast Cancer—Epidemiology, Risk Factors, Classification, Prognostic Markers, and Current Treatment Strategies—An Updated Review," by Lukasiewicz, Sergiusz, et al., 2021, Cancers, 13(17), p. 4287 (10.3390/cancers13174287). CC BY-NC. [87]

It is worth noting that having one or more of these risk factors does not necessarily mean a person will develop breast cancer [77]. Conversely, some women with breast cancer have no known risk factors. Early detection and regular screening are crucial for detecting breast cancer in its early stages when it is most treatable.

2.1.2 Treatments and Secondary Effects

The treatment for breast cancer depends on several factors, including the cancer stage, the type of breast cancer, and the patient's overall health. The most common treatments for breast cancer include:

- **Surgery** is typically the first treatment for breast cancer, where the tumor and a portion of surrounding healthy tissue are removed [54]. In some cases, a mastectomy may be necessary, which involves the removal of the entire breast.
- **Radiation therapy** is often used after surgery to destroy any remaining cancer cells and reduce the risk of recurrence [123].

- **Chemotherapy** is used when cancer has spread beyond the breast and uses drugs to kill cancer cells throughout the body [99].
- **Hormone therapy** blocks or lowers the level of certain hormones, such as estrogen, that can stimulate the growth of breast cancer cells [99].
- **Targeted therapy** uses drugs that target specific proteins or genes that promote the growth of cancer cells [55].

Each treatment may have side effects depending on the treatment and the patient. Common side effects of breast cancer treatment include fatigue, hair loss, nausea and vomiting, loss of energy, mouth sores, skin changes or rashes, menopausal symptoms (such as hot flashes, vaginal dryness, and mood swings), and lymphedema (swelling in the arm or hand) after surgery or radiation therapy [16].

It is essential for patients to discuss the potential side effects of their specific treatment plan with their healthcare provider and to seek support from healthcare professionals and support groups as needed[109]. The healthcare provider may recommend ways to manage the side effects and help the patient cope with breast cancer treatment's emotional and physical challenges.

Prognosis refers to the predicted outcome of a disease, including the likelihood of recovery and the expected duration and quality of life. The prognosis for breast cancer depends on various factors, including the stage of cancer at diagnosis, the type of breast cancer, the age and overall health of the patient, and the response to treatment [69].

Generally, the prognosis for breast cancer is better when the cancer is detected early and has not spread beyond the breast tissue. However, the prognosis may be less favorable if the cancer has spread to nearby lymph nodes or other body parts. The five-year survival rate for patients with early-stage breast cancer is over 90% [83].

2.1.3 Breast Cancer Screening

Breast cancer screening is a set of medical tests and exams to detect breast cancer before symptoms appear. Breast cancer screening aims to detect cancer early when it is more treatable and the chances of survival are higher. The most common breast cancer screening tests are mammograms, clinical breast exams, and breast self-exams [5]. Women with an increased risk of breast cancer may need to start screening earlier or have more frequent screenings. Women must discuss their risk factors and screening options with their healthcare providers. Many countries have developed their breast cancer screening guidelines based on population demographics, healthcare infrastructure, and resources, as no universal standard applies to all countries. While international organizations like the World Health Organization (WHO) and the International Agency for Research on Cancer (IARC) provide recommendations as a general framework for countries to adopt [106], the specific guidelines may differ between countries and organizations [105].

For instance, the American Cancer Society recommends annual mammograms for women at average risk of breast cancer starting at age 45 [127]. In contrast, the US Preventive Services Task Force recommends biennial mammograms for women aged 50 to 74. In the UK, the National Health Service suggests mammograms every three years for women aged 50 to 70

[35]. Mexico's breast cancer screening guidelines, as outlined in the NOM-041-SSA2-2011, recommend mammography every two years for women aged 40 to 69, annual clinical breast exams for women aged 25 to 39, and every two years for those aged 40 and over [33].

However, it is essential to remember that guidelines serve as general recommendations. Individuals must know their country's specific breast cancer screening guidelines and discuss their risk factors and screening options with their healthcare provider. Individual factors such as family history, genetic mutations, and personal preferences should also be considered when making decisions about breast cancer screening.

Mammography (low-dose X-ray imaging of the breasts) randomized controlled trials offer strong evidence that population screening substantially decreases breast cancer mortality by a relative risk of 20% for those invited to screening [64]. The efficacy of mammography screening relies on the age and is most noticeable in women aged 50-69, with less compelling evidence for benefits in other age groups [103]. Observational studies in real-world screening scenarios yield comparable evidence on the advantages of mammography screening to randomized controlled trials, though the estimated effects vary [81], [63]. Screening is known to enhance the early detection of breast cancer, leading to an anticipated benefit of reduced intensive treatments, such as lower mastectomy rates. Nevertheless, population-level research needs to present consistent findings regarding the impact of screening on treatment [131, 130].

2.2 Mammography

Mammography is a diagnostic imaging technique to detect abnormalities or changes in breast tissue [65]. It uses low-dose X-rays to create images of the breast, which can detect changes in breast tissue that may indicate breast cancer or other breast abnormalities. This technic is a widely used and effective screening tool for breast cancer. It can detect breast cancer in its early stages before it can be felt or seen on a physical exam [52]. Early detection is vital because it allows for more treatment options and better chances of survival [65].

Mammography can be performed using either analog film technology or digital technology [125]. Analog mammography uses X-ray film to produce images, while digital mammography uses electronic detectors to create images that can be viewed and stored on a computer [105]. Digital mammography offers several advantages over analog mammography, including faster image acquisition, better image quality, and the ability to manipulate and enhance images for better visualization [60]. Digital mammography also reduces the need for repeat exams and can detect more minor abnormalities than analog mammography [48].

There are two types of mammography exams [14]: screening and diagnostic mammography. Screening mammography is used to detect breast cancer in women with no disease symptoms. Diagnostic mammography evaluates abnormalities or changes in the breast tissue detected through screening mammography or other diagnostic tests, such as a breast MRI or ultrasound.

Mammography is generally recommended for women aged 40 to 74, although recommendations may vary depending on individual risk factors [85]. Women with a higher risk of breast cancer, such as those with a family history of the disease or a genetic mutation, may be advised to start screening at an earlier age or to have more frequent screening exams. Overall, mammography is considered the most effective screening tool for breast cancer, allowing for

the early detection of tumors and other abnormalities in the breast tissue [105].

2.2.1 Mammography Projections

Mammography can be performed in two primary projections: the craniocaudal (CC) projection and the mediolateral oblique (MLO) projection [124]. The CC projection is obtained by positioning the breast between two plates, with the X-ray tube above the breast and the image receptor below. [8] The CC projection views the breast from above and below, allowing a clear view of the entire breast tissue. The breast is compressed to obtain a clear image and reduce the amount of radiation exposure.

The MLO projection is obtained by positioning the breast at an angle between the X-ray tube, and the image receptor [8]. The breast is compressed similarly to the CC projection. In the MLO projection, the breast is viewed from the side, providing a clear view of the breast tissue from the chest wall to the nipple. This projection is particularly useful for detecting abnormalities in the upper outer quadrant of the breast, where many breast cancers are located. Please refer to Figure 2.1 for an example of mammography projections.

Additional projections may be performed in some instances, such as the lateral or spot compression views, to evaluate further suspicious areas detected in the CC or MLO projections.

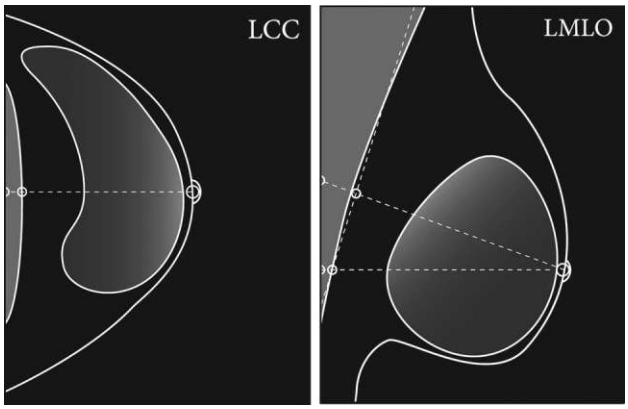

Figure 2.1: Mammographic Projections. Note. Adapted from "A review of mammo-graphic positioning image quality criteria for the craniocaudal projection," by Rhonda-Joy I Sweeney et al., 2018, The British journal of radiology, 91(1082), p. 20170611 (10.1259/bjr.20170611). CC BY-NC [133]

.

2.2.2 Mammography Interpretation and BI-RADS Score

Mammography interpretation is a multi-step process involving specialized training and expertise [12]. Typically, a radiologist specializing in breast imaging performs the interpretation. The process includes image evaluation, lesion detection, lesion characterization, and final assessment.

The first step in the process is image evaluation, where the radiologist reviews the mammogram images to ensure they are of sufficient quality for accurate interpretation [12];

This step includes checking for adequate breast compression, positioning, and exposure. If the images are not of sufficient quality, additional images may need to be taken. After ensuring image quality, the radiologist moves on to lesion detection. During this step, they carefully examine the mammogram for any signs of masses, calcifications, or other structural changes indicative of breast cancer or other conditions. If an abnormality is detected, the radiologist moves on to lesion characterization. This phase involves evaluating the lesion's size, shape, and other characteristics to determine the likelihood of cancerous abnormality.

BI-RADS

Based on the findings of lesion characterization, the radiologist assigns a final assessment category using the Breast Imaging Reporting and Data System (BI-RADS) [12]. The BI-RADS system is a standardized reporting system created by the American College of Radiology (ACR) to ensure consistency and accuracy in interpreting and reporting mammograms [5]. The BI-RADS score ranges from 0 to 6, with each number corresponding to a specific level of suspicion and management recommendations.

A score of 0 indicates that additional imaging or evaluation is needed to determine the presence of a breast abnormality [129]. A score of 1 is assigned when the mammogram is regular with no evidence of breast abnormalities. A score of 2 means that there is a benign finding, such as a cyst, that does not require further evaluation. A score of 3 indicates a benign finding that requires a short-term follow-up mammogram to confirm stability or resolution. A score of 4 suggests a suspicious abnormality that may require a biopsy, and it is further classified into subcategories based on the level of suspicion. A score of 5 indicates a highly suspicious abnormality, likely malignant, and requires immediate biopsy. Finally, a score of 6 is assigned when there is a known biopsy-proven malignancy.

The BI-RADS score helps standardize the interpretation and reporting of mammograms and provides a consistent language for communication between healthcare providers [34]. It ensures that suspicious findings are appropriately evaluated and managed, reducing the risk of missed diagnoses or unnecessary procedures.

2.2.3 Additional Diagnostic Imaging Techniques and Advantages of Mammograms

In addition to mammography, several other diagnostic imaging techniques may help diagnose breast cancer and other conditions [65]. Several of these strategies include [49]:

- Ultrasound: A breast ultrasound produces images of the breast tissue using high-frequency sound pulses. It is frequently used with mammography to evaluate suspicious findings further or examine breast tissue in women with dense breasts.
- Magnetic resonance imaging (MRI): This technic creates detailed images of breast tissue using a magnetic field and radio waves. In conjunction with mammography, it is frequently used to evaluate further suspicious findings in women at high risk for breast cancer.
- Tomosynthesis: This technic, also known as 3D mammography, is a sophisticated form of mammography that produces a three-dimensional image of breast tissue. It can be

used with conventional mammography to produce more detailed images and enhance diagnostic precision.

The choice of additional imaging techniques will hinge on several factors, including the woman's age, breast density, and the nature of the mammographic findings [65]. The objective is to use the most suitable imaging or combination of techniques to obtain an accurate diagnosis and direct the most appropriate treatment.

Mammography is superior to other diagnostic imaging methods in several ways [105]. First, because it produces high-quality images of the breast tissue while only using small amounts of radiation, it is a non-invasive and relatively low-risk technique. As a result, it can be used as a reliable screening technique to find breast cancer in its early stages.

Second, because mammography is common in most healthcare facilities, it is generally accessible and available [105]. Since it is a very accessible method, it makes it possible to test for breast cancer quickly and easily, essential for improving outcomes and raising survival rates.

Thirdly, mammography is a recognized imaging method that has undergone extensive research and development [49]. As a result, standardized processes and reporting systems, such as the BIRADS score, have been developed to guarantee uniformity and precision in interpretation and reporting.

The effectiveness of mammography in lowering breast cancer mortality rates has also been demonstrated [17]. It becomes a vital weapon in the fight against breast cancer and in enhancing women's health outcomes.

2.3 Breast Cancer Risk Assessment Tools

Breast cancer prediction models are crucial for estimating a woman's risk of developing breast cancer [123]. These models offer valuable information to both women and healthcare providers using factors such as personal and family medical history, lifestyle, and genetics. In addition, they can identify women who may require more screening or preventive measures, enabling early detection and management of the disease. With advancements in technology, these models have become more comprehensive and precise.

The models fall into two main categories [19]: statistical-based models and image-based models. Statistical-based models use factors like age, family history, and breast density to calculate risk, while image-based models use imaging to derive a risk index. Well-known models include the Gail Model (Breast Cancer Risk Assessment Tool), the Tyrer-Cuzick Model, and the Claus model. As these models continue to evolve, they are expected to improve further, potentially leading to better outcomes for women at risk of breast cancer.

2.3.1 Statistical Based-Models

The Gail Model

The Gail Model, or the Breast Cancer Risk Assessment Tool, is a scientific tool created by Dr. Mitchell Gail and his team in the late 1980s [44]. Its primary purpose is to estimate a woman's likelihood of developing severe breast cancer within the next five years or by age 90. The model helps healthcare professionals identify women at a higher risk of developing breast

cancer, allowing doctors to recommend additional health checks, preventative measures, or genetic counseling as needed.

However, the Gail Model does have its limitations. While it considers seven essential factors [44] like age, family history of breast cancer, and race/ethnicity, it does not account for lifestyle factors such as alcohol use, physical activity, or weight. Its accuracy may be less for certain types of breast cancer [150], and it was primarily developed using data from Caucasian women, potentially limiting its accuracy for other racial and ethnic groups [27].

The Tyrer-Cuzick model

The Tyrer-Cuzick Model, also known as the International Breast Cancer Intervention Study (IBIS) model, is a statistical tool developed in the early 2000s by Jack Cuzick and his team to estimate a woman's risk of developing breast cancer over specific periods [140], such as the next ten years. Compared to models like the Gail Model, it provides a more comprehensive risk assessment, aiding healthcare providers in identifying women at high risk for breast cancer [61]. This identification contributes to more personalized risk assessment and prevention measures, including additional screening, preventive measures, or genetic counseling.

However, the Tyrer-Cuzick Model has some limitations. For example, it may overestimate the risk in certain groups, such as women with a family history of breast cancer [80]. It also does not fully consider lifestyle factors like alcohol consumption or physical activity, which can influence breast cancer risk. In addition, predominantly based on data from UK women, its accuracy for other racial and ethnic groups might be limited. Lastly, it can only estimate risk, not predict with certainty whether an individual will develop breast cancer.

The Claus Model

Developed by Dr. Ephraim Paul Claus in the early 1990s, the Claus Model is a statistical tool used to estimate a woman's risk of developing breast cancer based on her family history [28]. This model helps healthcare providers assess breast cancer risk in women with a significant family history, aiding in personalized risk assessment and prevention strategies. The Claus Model's main strength is its focus on a family history of breast cancer. It considers the number of affected first-degree relatives and their ages at diagnosis to provide a tailored understanding of a woman's potential risk.

However, the Claus Model has several limitations [39]. It cannot predict with certainty if an individual will develop breast cancer but only estimates risk. It primarily focuses on family history, ignoring other crucial risk factors such as age, reproductive history, and lifestyle factors. Furthermore, its effectiveness may vary across different populations, as it primarily uses data from Caucasian women in the U.S. Also, the model might not be as widely used or accessible as other models like the Gail or Tyrer-Cuzick models, limiting its use in different contexts.

2.3.2 Image-Based Models

On the other hand, the image-based models evaluate the features presented in the mammograms [1]. Furthermore, these methods implement machine learning and deep learning techniques to predict the risk of developing Breast Cancer. In addition, there are metrics to evaluate the performance of a machine learning algorithm [27].

In this research, it is proposed the implementation of machine learning techniques to obtain a risk assessment model. Every risk assessment tool is validated through discrimination

and calibration scores. Discrimination evaluates the ability of the model to separate illness and healthy samples, where 1.0 means perfect qualification and 0.5 no discrimination performed [27]. In addition, the tool's performance is validated for calibration in the total sample population (observed-to-expected ratio).

The current work in Breast cancer risk prediction showed a poor performance [86]. The models developed have internal tooling, infrastructure, and hardware limitations, leading to no practical approaches [50]. It is intended to design a model with increased discrimination and calibration parameters. In addition, in the work of Pal Choudhury et al., [26], it is mentioned that the improvements and adjustments of the already developed risk prediction tools are not accurate, and it is complicated to use them in different racial and ethnic populations. Therefore, the current challenge is designing a model that can perform well in a varied population.

2.4 Pre-processing of Images

Pre-processing is a crucial step in image analysis that involves various techniques to clean, enhance, and transform images [128]. It generally includes several steps, such as image segmentation, normalization, noise reduction, and image registration. Image segmentation separates different regions within an image [111], and normalization adjusts the brightness and contrast of an image for improved quality and consistency [158]. Additionally, image denoising can eliminate unwanted noise [128], while image registration aligns images to a standard coordinate system. By enhancing the quality of images and reducing artifacts, pre-processing can optimize the accuracy and efficiency of image analysis tasks, including feature extraction [23].

2.4.1 Image Segmentation

Image segmentation is a technique that divides an image into distinct segments based on specific characteristics like color, texture, or intensity [111]. It aims to identify and isolate different objects or regions of interest within an image for further analysis. Image segmentation is crucial in diagnosing and treating various diseases in medical imaging [23]. Medical images often contain complex structures and variations in tissue densities, making it challenging to analyze specific regions accurately. Image segmentation techniques help overcome these challenges by identifying and isolating structures like tumors, blood vessels, or anatomical features [154].

Thresholding is a commonly used approach in image segmentation that creates a binary partition of an image based on its intensities [154]. The goal is to determine a threshold value that separates desired classes within the image. This process involves identifying peaks and valleys in the image's histogram, where valleys represent potential threshold values. Multithresholding can also be used to determine multiple thresholds. Thresholding is a straightforward and effective method for segmenting images with contrasting intensities or other quantifiable features. While interactive methods are often employed, automated techniques also exist [118]. Thresholding is frequently used as an initial step in image processing, particularly in applications like digital mammography, where different tissue classes, such as healthy and tumorous, must be distinguished [111].

In conclusion, image segmentation is a vital technique in medical imaging that helps identify and isolate specific regions or structures of interest within complex medical images. Thresholding, as a commonly used approach in image segmentation, enables the creation of a binary partition based on image intensities, allowing for the separation of desired classes. This straightforward method effectively segments images with contrasting intensities and plays an important role in various image-processing applications, including digital mammography. Both interactive and automated thresholding techniques contribute to improved analysis and diagnosis in medical imaging, ultimately leading to better treatment outcomes [154, 23, 111, 118].

2.4.2 Image Normalizatiom

It is necessary for medical imaging to ensure consistent and comparable images across patients, scanners, and imaging protocols [142]. Image normalization is a crucial pre-processing technique to enhance image quality and consistency by adjusting brightness and contrast [58]. By standardizing the intensity range, image normalization eliminates variations in luminance and contrast, enabling more reliable analysis.

Moreover, it plays a vital role in feature extraction, improving the accuracy and reliability of identifying and extracting pertinent image information. Standard image normalization techniques include linear scaling, histogram equalization, and contrast stretching [51]. Linear scaling is commonly used in grayscale images to scale pixel intensities within a specific range. Histogram equalization redistributes pixel intensities to achieve a uniform distribution and enhance image contrast. Contrast stretching stretches intensity values to increase image contrast.

Image normalization is essential for various image-processing applications, especially in medical imaging [92]. It ensures comparability and uniformity of images, enhances feature extraction accuracy, and improves analysis reliability by minimizing differences in brightness and contrast levels.

2.4.3 Wavelet Decomposition

Image wavelet decomposition involves decomposing an image into its frequency components using wavelets [100]. At each level, the image is divided into four sub-bands: approximation, horizontal detail, vertical detail, and diagonal detail [110]. Four images are obtained from each level. Wavelet decomposition of mammographies helps isolate and analyze different structures in breast tissue, aiding in detecting and diagnosing abnormalities [144].

The Haar wavelet stands out among other mother wavelets for image wavelet decomposition [108]. It is simple, computationally efficient, and effectively captures abrupt changes and edges. In addition, the Haar wavelet's balanced frequency response and correct localization in spatial and frequency domains make it suitable for image compression, denoising, and feature extraction.

2.5 Feature Extraction

Feature extraction is crucial in data analysis [113], particularly in complex and high-dimensional fields like image analysis and machine learning. It plays a significant role in transforming raw data into a more suitable format for analysis or modeling. By identifying and selecting the most relevant features, feature extraction reduces the dimensionality of the data [147], improving analysis efficiency and accuracy. It also helps mitigate the risk of overfitting in models, where performance on new data can be poor despite good performance on training data.

In image analysis, feature extraction involves identifying and extracting relevant information such as texture, color, and shape from digital images [155]. These features can be used to train machine learning models for tasks like image classification, object recognition, and image segmentation. In machine learning, feature extraction helps improve model performance by selecting the most critical features as algorithm inputs. This technique is valuable in high-dimensional and complex datasets, such as medical image processing.

2.5.1 Image Feature Extraction

In image feature extraction, first-order and second-order features refer to different statistical properties that can be derived from an image [4].

First-order features, also known as statistical features or histogram-based features, are computed directly from the pixel intensities of the image [91]. These features provide basic statistical information about the distribution of pixel values within the image. Common first-order features include:

- Mean: The average pixel intensity of the image.
- Standard Deviation: The measure of the spread or variability of pixel intensities.
- Skewness: A measure of the asymmetry of the pixel intensity distribution.
- Kurtosis: A measure of the peakedness or flatness of the pixel intensity distribution.
- Entropy: A measure of the randomness or uncertainty of the pixel intensities.

On the other hand, second-order features, also known as texture features or spatial features, capture the spatial relationships between pixels [91]. These features are derived from the gray-level co-occurrence matrix (GLCM) or the gray-level run-length matrix (GLRLM), which analyze pixel pairs or runs of pixels in the image. Second-order features describe the texture or pattern characteristics of the image. Some commonly used second-order features include:

- Contrast: Measures the local intensity variations between neighboring pixels.
- Energy: Represents the sum of squared pixel intensity values in the image.
- Homogeneity: Measures the image's similarity or uniformity of pixel intensities.
- Correlation: Indicates the linear dependency between pixel intensities.

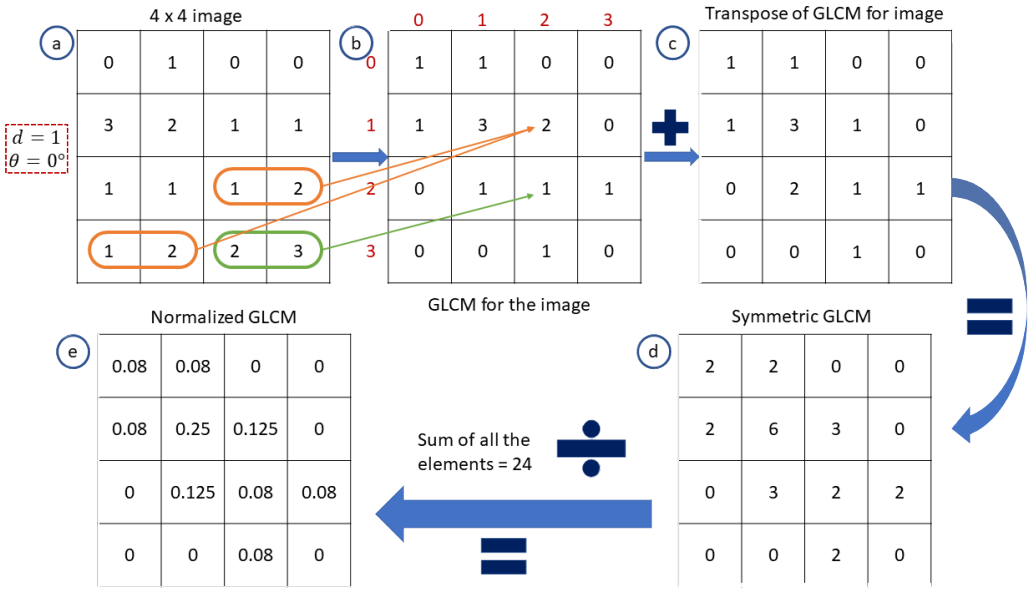

Figure 2.2: Calculation of Gray Level Co-ocurrence Matrix of a 4 × 4 image for distance $d = 1$ and direction $\theta = 0$.

- Dissimilarity: Measures the differences in pixel intensities between neighboring pixels.

These first-order and second-order features provide valuable information about the image content and characteristics, and they are widely used in various image processing and computer vision applications [4].

2.5.2 Gray-Level Co-occurrence Matrix Features

Gray-level co-occurrence matrix (GLCM) features are an image analysis technique for extracting texture features [57]. GLCM features describe the spatial relationship between pairings of pixels with the same gray-level intensity [121]. They are particularly beneficial for the texture characterization analysis of medical images.

GLCM features are computed by constructing a matrix that tallies the number of occurrences of a pair of pixels with the same gray-level intensity in a particular spatial relationship [57]. The distance and angle between them can define the spatial relationship between pixels (see figure 2.2 for reference). After constructing the GLCM, a set of statistical features can be derived from the matrix. The most prevalent GLCM image analysis features include the following [4]:

Contrast measures the differences in pixel intensities between adjacent pixels. It is calculated as the sum of the squared differences in gray-level values.

Energy: Energy, known as homogeneity or angular second moment, quantifies the texture homogeneity. It is computed as the sum of the squared GLCM elements.

Entropy measures the randomness or uncertainty of an image's texture. It is computed by adding the products of the non-zero GLCM elements and their logarithms.

Homogeneity: Homogeneity evaluates the proximity of the GLCM element distribution to the diagonal. It is determined by dividing the total number of elements in the GLCM by the

absolute difference between the row and column indices.

Correlation measures the linear relationship between the gray-level values of adjacent pixels. It is determined by adding the products of the row and column indices in the GLCM, subtracting the mean, and dividing by the standard deviation.

2.5.3 Local Binary Pattern Features

Local Binary Pattern (LBP) is an image processing and computer vision technique for texture analysis [68]. LBP features characterize the local patterns of an image by comparing the gray level of each pixel to the gray level of its neighbors.

LBP features are computed by comparing each pixel's gray level to its neighboring pixels' gray levels [15]. Each pixel is assigned a binary code based on whether the neighboring pixel values exceed the central pixel's value. These binary codes are then used to construct an image-wide histogram of LBP patterns. The histogram can represent the image's texture as a feature vector. The LBP (Local Binary Pattern) code is defined by Equation 2.1 when considering a center pixel (X_c) and its neighboring pixels within a radius (R), denoted by P [15].

$$LBP_{P,R} = \sum_{i=0}^{P-1} s(X_i - X_c) \cdot 2^i, \tag{2.1}$$

$$s(X_i - X_c) = \begin{cases} 1 & \text{if } X_i - X_c \geq 0 \\ 0 & \text{if } X_i - X_c < 0. \end{cases} \tag{2.2}$$

The calculation of LBP involves the following steps [15]:

1. Choose a window and identify the neighbors of the center pixel X_c based on the radius R. When R equals 1, there will be eight neighbors, and when R is equal to 2, there will be 16 neighbors.

2. Compute the binary bit value for each neighbor by utilizing Equation 2.2.

3. Multiply each binary bit value with the corresponding weight mask value.

4. Calculate the sum.

Figure 2.3 illustrates the procedure of the algorithm using a diagram, showcasing the various steps involved.
As a result, Local Binary Pattern (LBP) finds extensive application in tasks such as image matching, the detection, and tracking of pedestrians and car targets, and the analysis of biological and medical images [112].

2.5.4 Radiomics

Radiomics, a new field in medical imaging, involves extracting and analyzing quantitative features from medical images to uncover connections between these features and clinical outcomes [22]. This practice could revolutionize personalized medicine by offering enhanced insights into disease diagnosis, prognosis, and treatment response. Challenges to address

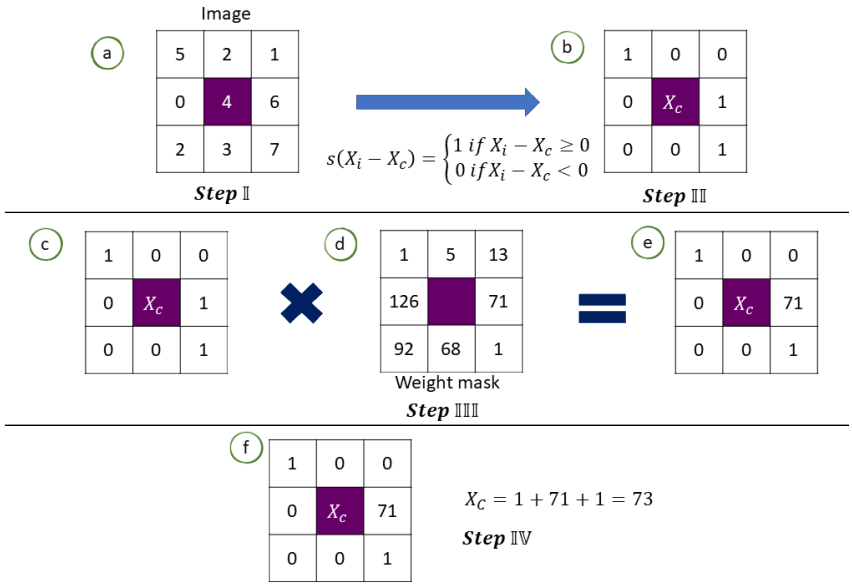

Figure 2.3: In Panel (A), the neighboring pixels of the center pixel ($X_c = 4$) within a radius of $R = 1$ are depicted. Panel (B) displays the binary values computed for each neighbor of X_c using Equation 2.2. Panels (C-E) illustrate the evaluation of step 3 in the LBP algorithm, with Panel (D) representing the weight mask. Lastly, Panel (F) presents the LBP value of the center pixel X_c.

include ensuring reproducibility across different imaging platforms and evaluating feature stability [90].

The field faces challenges such as overfitting due to the high number of extractable features, which can be mitigated by feature selection and dimensionality reduction [79]. Efforts are ongoing to develop robust and clinically relevant radiomic signatures to enhance disease diagnosis and treatment.

Radiomic features should be biologically relevant and interpretable within disease biology [90]. Validation of these features and signatures is crucial in diverse cohorts to ensure their clinical utility. Integrating radiomic features with other clinical and imaging data is critical for practical clinical application. Current research focuses on using radiomics to understand disease mechanisms further and enhance patient stratification and treatment selection [3].

2.6 Feature Engineering

Feature engineering is essential in machine learning, transforming raw data into valuable features to increase the predictive accuracy of algorithms [157]. This process involves selecting, extracting, and transforming data into more informative, problem-relevant features. The quality of these features directly affects the model's performance, with well-designed features allowing the model to identify data patterns effectively, improve its performance, and prevent overfitting [119].

Feature engineering has wide-ranging applications such as text analysis, image recognition, and time-series forecasting [21]. Techniques used in feature engineering include feature selection, extraction, and transformation, aiming to pinpoint significant features, create new ones through mathematical or statistical methods, and adjust data to fit the machine learning algorithm [143].

2.6.1 Univariate Analysis

In the realm of feature extraction, univariate analysis refers to the process of evaluating one feature (variable) at a time [134]. This type of analysis helps understand each feature's properties and statistical characteristics, independent of any other variables.

The primary aim of univariate analysis in feature extraction is to assess each feature's potential for contributing valuable information to a machine learning model or data analysis process [120]. The characteristics usually examined include mean, median, mode, standard deviation, variance, range, skewness, kurtosis, and frequency distribution of the data, along with checking for outliers.

Univariate analysis is essential in data analysis and machine learning [67]. It helps understand the data by identifying patterns, outliers, and trends. It also assists in feature selection and preprocessing decisions and visualizes feature properties. Univariate analysis is crucial in making informed choices for feature extraction and model building in machine learning.

2.7 Cox Models

The Cox Proportional Hazards (CoxPH) model is a prominent tool for analyzing time-to-failure data [11]. Established by Cox in 1972 [32], it is a commonly used regression method in medical research to identify the relationship between a patient's survival time and one or more predictor variables [11, 138]. The model also allows calculating an event's hazard (risk) based on individual characteristics or prognostic variables.

At its core, the CoxPH model enables an examination of how certain factors impact the rate of events, such as surgery, death, or changes in medical condition, at any given time [11]. This inference is often referred to as the risk rate. The model employs a hazard function, h(t), to represent the risk of an event happening at a particular time based on a set of variables and their coefficients [138].

This model is deemed semi-parametric [73], as it includes an unspecified baseline hazard function, h0(t), representing the hazard if all variables equal zero. The model's strengths lie in its ability to assume that the hazard of an event remains constant over time, hence the term "proportional hazards".

The model also provides a Hazard Ratio (HR) for each variable, representing its effect on risk [73]. If the coefficient is more significant than zero, the risk increases, event time decreases with an increase in the variable value, and vice versa. Despite not specifying the baseline hazard, the CoxPH model is renowned for providing accurate estimations and risk indexes, making it a robust and highly favored tool in many data situations.

2.8 Machine Learning

Machine learning is a subset of artificial intelligence that involves the development of algorithms that can learn from and make predictions or decisions based on data [89]. These algorithms improve their performance as the amount of data they are exposed to increases. Machine learning is used in various applications, including natural language processing, image recognition, and recommendation systems [145]. It can transform many industries by enabling computers to automatically identify patterns and make data-based decisions.

2.8.1 BSWiMS

BSWiMS (Bootstrapped Stage-wise Model Selection) is a supervised model selection method that aims to construct the most accurate predictive model using Cox models [107]. The model comprises 'nuggets' or compact linear models built from a unique set of statistically significant characteristics [136]. The BSWiMS process has five stages [107]:

1. **Univariate Filter:** This stage uses the Benjamini-Hochberg procedure to select features with a robust univariate association with the outcome.
2. **Bootstrapped Forward Selection:** This phase involves building numerous linear models using forward selection, each introducing additional significant features until no further improvement is seen.
3. **Frequency-based Forward Selection:** Here, features are ordered by frequency in bootstrapped models to generate a single model.
4. **Backward Elimination:** Non-significant features are removed from the model using bootstrapping, resulting in a nugget-model where all features are statistically significant. This step is repeated until no further models can be found or the performance decreases.
5. **Model Bagging:** This final stage involves aggregating all the nugget-models into a single statistical model.

The resulting BSWiMS model is unique, with features selected based on average nugget-fitted fitness statistics and selection frequency [136].

2.8.2 LASSO

Least Absolute Shrinkage and Selection Operator (LASSO), a regression analysis method, combines variable selection and regularization to improve prediction accuracy and model interpretability [138]. The LASSO regression model introduces a penalty term discouraging the sum of absolute coefficients. Increasing the value of λ intensifies the penalty strength, resulting in a more significant shrinkage of coefficients and eliminating additional features from

the model [59]. This regularization technique addresses the issue of overfitting commonly observed in linear regression. Equation 2.3 presents the definition of LASSO regression [59].

$$\min_\beta \frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - ... - \beta_p x_{ip})^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{2.3}$$

where:

$$\begin{aligned} y_i &= \text{the } i\text{th observation of the dependent variable} \\ x_{ij} &= \text{the } i\text{th observation of the } j\text{th independent variable} \\ \beta_0 &= \text{the intercept} \\ \beta_j &= \text{the coefficient for the } j\text{th independent variable} \\ \lambda &= \text{the regularization parameter.} \end{aligned}$$

The steps involved in LASSO regression are as follows:

1. Split the data into training and testing sets to train and evaluate the model.
2. Choose a value for the lasso penalty, which determines the amount of regularization. A higher penalty leads to more regularization and fewer features in the model.
3. Fit the model to the training set using gradient or coordinate descent methods.
4. Evaluate the model's performance on the testing set using metrics such as AUC or accuracy.
5. Select the best model by comparing the performance of different models on the testing set.

LASSO regression balances bias and variance by dividing the data and selecting an appropriate penalty. The splitting helps prevent overfitting and improves the model's generalization ability to new data [59].

2.8.3 Cross-Validation

Cross-validation is a widely used technique in machine learning for assessing a model's performance and generalization ability [75, 72]. It involves dividing the dataset into subsets and training the model on some subsets while evaluating it on others. The most common form is k-fold cross-validation [115], where the dataset is split into k equal-sized folds, and the model is trained and evaluated multiple times. Cross-validation provides a more robust performance estimate, helps detect overfitting or underfitting, and aids in selecting optimal hyperparameters. It is a valuable tool for researchers and practitioners to obtain reliable evaluations and gain insights into a model's capabilities.

Leave-One-Out Cross Validation

Leave-one-out cross-validation (LOOCV) evaluates a machine learning model by training the model on all but one data point and then testing the model on the remaining data point [20].

This process is repeated for each data point in the dataset. The average error across all data points is then used to evaluate the model's performance.

The LOOCV method is suitable for small sample sizes, addressing potential issues encountered by other techniques like tenfold or fivefold cross-validation [47]. LOOCV is a computationally expensive method but is also one of the most accurate methods. The reason for the computational cost is that LOOCV uses all of the data in the dataset to evaluate the model, which helps to reduce the bias in the evaluation.

2.9 Machine Learning Algorithms Evaluation

In machine learning, metrics are used to evaluate the performance of a model. One commonly used metric is the confusion matrix, which provides a breakdown of true positives, true negatives, false positives, and false negatives. Several other metrics can be derived from the confusion matrix, including accuracy, sensitivity, specificity, and false positive rate. Accuracy measures the proportion of correct predictions made by the model, while sensitivity and specificity measure the proportion of actual positive and negative cases correctly identified by the model. The false positive rate is the proportion of actual negative cases incorrectly identified as positive by the model. Another widely used metric is the Receiver Operating Characteristic (ROC) curve and its corresponding area under the curve (AUC), which measures the model's ability to discriminate between positive and negative cases.

2.9.1 Confusion Matrix

The confusion matrix is a table commonly used to describe the performance of a classification model on a set of test data [56]. It allows us to visualize a classification model's correct and incorrect predictions (see figure 2.4). The matrix contains four values: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). The diagonal elements represent the number of correct predictions, and the off-diagonal elements represent the number of incorrect predictions.

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Figure 2.4: Confusion Matrix of a Machine Learning Model

- True Positives: Samples ***properly*** classified as positive.
- False Positives: Samples ***misclassified*** as positives.
- False Negatives: Samples ***misclassified*** as negatives.
- True Negatives: Samples ***properly*** classified negatives.

The confusion matrix can obtain Several essential metrics, including accuracy, sensitivity, specificity, and false positive rate.

Accuracy

Accuracy is the proportion of correct predictions over the total number of predictions made [94]; please see equation 2.4 for reference. It measures the model's ability to classify positive and negative instances correctly.

$$Accuracy = \frac{Correct\ Predictions}{Total\ Predictions}. \tag{2.4}$$

Sensitivity

Sensitivity, also known as recall, measures the proportion of true positive predictions over the total number of actual positive instances; you can find equation 2.5 as a reference. It indicates how well the model performs in identifying the positive class [62].

$$Sensitivity = \frac{True\ Positives}{False\ Negatives + True\ Positives}. \tag{2.5}$$

Specifity

Specificity measures the proportion of true negative predictions over the total number of actual negative instances, consult equation 2.6. It indicates how well the model is performing in identifying the negative class [62].

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives}. \tag{2.6}$$

False Positive Rate

The false positive rate (FPR) is the proportion of false positives over the total number of actual negatives; please see equation 2.7 for reference. It indicates how often the model needs to be corrected when it predicts the positive class [62].

$$FalsePositiveRate = 1 - Specificity. \tag{2.7}$$

The confusion matrix and the metrics derived from it can provide insights into the strengths and weaknesses of a classification model, helping to identify areas for improvement and further tuning.

2.9.2 ROC AUC

The Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) are commonly used to evaluate the performance of binary classification models [43]. The ROC curve represents a graphical trade-off between the true positive rate (Sensitivity) and the false

positive rate (1 - Specificity) at various classification thresholds. The AUC is a metric that quantifies the model's overall performance based on the ROC curve. This graphic is obtained as shown in figure 2.9.2.

The ROC curve is generated by varying the classification threshold of a model and calculating

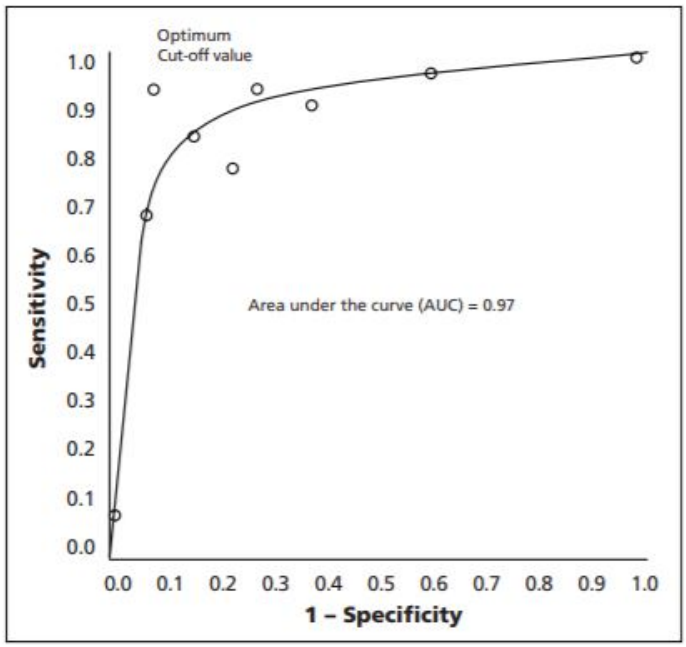

Figure 2.5: ROC curve illustrating high discriminatory power. Note. Adapted from "Understanding receiver operating characteristic (ROC) curves" by Fan, J., Upadhye, S., & Worster, A., 2006, Canadian Journal of Emergency Medicine, 8(1), 19-20. [41]

Sensitivity and $(1 - Specificity)$ values [114]. It represents different thresholds, with the ideal curve close to the top-left corner, indicating high sensitivity and a low false positive rate. The AUC (Area Under the Curve) is a measure ranging from 0 to 1, where a higher value indicates a better ability to distinguish between positive and negative instances. The ROC AUC is particularly useful in imbalanced datasets, offering a comprehensive evaluation of the model's performance by considering both Sensitivity and Specificity, overcoming limitations of accuracy when one class dominates the results.

2.10 Related Works

In recent years, integrating radiomics and machine learning techniques has emerged as a promising approach for predicting the risk of developing breast cancer. Numerous studies have investigated the potential of utilizing quantitative imaging features extracted from medical images and advanced machine learning algorithms to enhance breast cancer risk assessment [97]. These works have leveraged various imaging modalities, such as mammography, magnetic resonance imaging (MRI), and ultrasound, to capture detailed information about tumor characteristics and tissue composition. By analyzing these radiomic features and employing

powerful machine learning models, researchers have aimed to develop accurate and reliable predictive models to aid in early detection and personalized treatment planning for breast cancer patients. This comprehensive review presents an overview of the essential findings and methodologies employed in implementing radiomics and machine learning techniques for breast cancer risk prediction, highlighting the significant progress made in this rapidly evolving field.

The work of Eriksson et al. presents the results of a case-cohort study of 8,604 randomly selected women within a mammography screening cohort initiated in 2010 in Sweden for women aged 40-74 years [37]. Two thousand twenty-eight incident breast cancers were identified through register matching in May 2022 (206 incident breast cancers were found in the subcohort). Mammograms, age, lifestyle, and familial risk factors were collected at the study entry.

The Tyrer-Cuzick v8 risk model incorporates self-reported lifestyle and familial risk factors and mammographic density to estimate risk [37]. The image-based model extracted mammographic features (density, microcalcifications, masses, and left-right breast asymmetries of these features) and age from study entry mammograms. Absolute risks were estimated, and age-adjusted AUC model performances (aAUCs) were compared across ten years. The age-adjusted AUCs (aAUCs) of the image-based risk model ranged from 0.74 (95% CI, 0.70 to 0.78) to 0.65 (95% CI, 0.63 to 0.66) for breast cancers developed 1-10 years after study entry; the corresponding Tyrer-Cuzick aAUCs were 0.62 (95% CI, 0.56 to 0.67) to 0.60 (95% CI, 0.58 to 0.61). For symptomatic cancers, the aAUCs for the image-based model were $\geq$0.75 during the first three years.

A similar work to the previous one presents the development of a mammography-based deep learning (DL) breast cancer risk model that is more accurate than established clinical breast cancer risk models [151].

The study included 88,994 consecutive screening mammograms in 39,571 women between January 1, 2009, and December 31, 2012. Each patient's examinations were assigned to training, validation, or test sets. Cancer outcomes were obtained through linkage to a regional tumor registry. By using risk factor information from patient questionnaires and electronic medical records review, three models were developed to assess breast cancer risk within five years: a risk-factor-based logistic regression model (RF-LR) that used traditional risk factors, a DL model (image-only DL) that used mammograms alone, and a hybrid DL model that used both traditional risk factors and mammograms [151].

The hybrid DL model showed significantly higher AUCs than the Tyrer-Cuzick model (version 8) and RF-LR. The study concluded that deep learning models that use full-field mammograms yield substantially improved risk discrimination compared with the Tyrer-Cuzick (version 8) model [151].

Another work developed presents the development of a mammography-based deep learning model called "Mirai" designed to predict breast cancer risk at multiple time points, leverage potentially missing risk factor information, and produce predictions that are consistent across mammography machines [152].

Mirai was trained on a large dataset from Massachusetts General Hospital (MGH) in the United States and tested on held-out test sets from MGH, Karolinska University Hospital in Sweden, and Chang Gung Memorial Hospital (CGMH) in Taiwan. The model obtained a 5-year AUC of ¡0.7, significantly higher than the Tyrer-Cuzick model and prior deep learning

models, Hybrid DL, and Image-Only DL, trained on the same dataset. Mirai more accurately identified high-risk patients than prior methods across all datasets [152].

It has been demonstrated that the use of Machine learning techniques and radiomics information has been applied in the medical field. Different applications, such as disease diagnosis, performed well using Machine learning. In the work of Tamez et al., [137], ML and radiomics information were applied to help diagnose Breast Cancer in digital mammograms. Research has included machine learning techniques to predict the survival of Breast Cancer Patients [84].

The work of Kontos et al. displays a study that aimed to identify phenotypes of mammographic parenchymal complexity by using radiomic features and to evaluate their associations with breast density and other breast cancer risk factors [76].

The study used computerized image analysis to quantify breast density and extract parenchymal texture features in a cross-sectional sample of women screened with digital mammography. Differences across phenotypes by age, body mass index, breast density, and estimated breast cancer risk were assessed. Unsupervised clustering was applied to identify and reproduce phenotypes of parenchymal complexity in separate training and test sets. The study concluded that radiomic phenotypes might augment breast cancer risk prediction models [76].

Another study illustrates the development and internal validation of a digital breast tomosynthesis (DBT)-based short-term risk model for predicting future late-stage and interval breast cancers after negative screening exams [38].

The study included 805 incident breast cancers and a random sample of 5173 healthy women matched on the year of study entry in a nested case-control study from 154,200 multiethnic women aged 35 to 74 attending DBT screening in the United States between 2014 and 2019 [38]. A relative risk model was trained using elastic net logistic regression and nested cross-validation to estimate risks for using imaging features and age. An absolute risk model was developed using derived risks, U.S. incidence, and competing mortality rates.

The discrimination performance of 1-year risk was 0.82 (95% CI, 0.79 to 0.85) with good calibration (P = 0.7). Using the U.S. Preventive Service Task Force guidelines, 14% of the women were at high risk, 19.6 times higher than the general risk. In this high-risk group, 76% of stage II and III cancers and 59% of stage 0 cancers were observed (P ¡ 0.01).

The work developed by Gastounioti et al. explains an external validation on a U.S. screening cohort of a mammography-derived AI breast cancer risk model initially developed for European screening cohorts [46].

The study retrospectively identified 176 breast cancers with exams three months to 2 years before cancer diagnosis and a random sample of 4963 controls from women with at least one year of negative follow-up. The AI risk model showed a discriminatory performance of AUC 0.68, comparable to previously reported European validation results (AUC = 0.73). The discriminatory performance of the AI risk model was non-significantly different by race (AUC for White women = 0.67 and Black women = 0.70), p = 0.20 [46].

Concerning a clinically used lifestyle–family-based risk model, the AI risk model showed a significantly higher discriminatory performance (AUCs 0.68 vs. 0.55, p ¡ 0.01) [46].

The study concluded that using mammographic features generated from DBT screens, their image-based risk prediction model could guide radiologists in selecting women for clinical care, potentially leading to earlier detection and improved prognoses [38].

The article by Lee et al. shows a new method called PRIME+ for breast cancer risk prediction that leverages prior mammograms using a transformer decoder, outperforming a state-of-the-art risk prediction method that only uses mammograms from a single time point [82].

The study validated their approach on a dataset with 16,113 exams. Experimental results show that their model achieves a statistically significant improvement in performance over the state-of-the-art-based model, with a C-index increase from 0.68 to 0.73 (p ¡ 0.05) on held-out test sets. Further, it demonstrated that it effectively captures patterns of changes from prior mammograms, such as changes in breast density, resulting in improved short-term and long-term breast cancer risk prediction [82].

Additionally, several improvements have been made in enhancing a risk assessment model for Breast Cancer. As the problem definition mentions, some evaluate the non-image and image risk factors through statistical techniques. To improve the accuracy to predict the risk of developing Breast Cancer, scientists started using Machine Learning techniques. A work that uses non-image factors in an ML algorithm was performed by Asri et al., which evaluates different supervised learning classifiers [10]. Using numerical attributes with Machine Learning and Random Optimization for Breast Cancer Prognosis showed a good performance [42]. Other work that implements Machine Learning in cancer prognosis and the prediction was performed by Kourou et al. [78].

2.11 Summary

This chapter provides a detailed literature review of this thesis, covering various aspects of breast cancer. Firstly, it discusses the background of breast cancer, including historical statistics and an overview of diagnosis and screening protocols. Additionally, it introduces some clinical tools used for risk assessment in this disease.

In addition to exploring topics related to breast cancer, the chapter also delves into concepts relevant to the proposed methodology. Some relevant definitions include a discussion on pre-processing image techniques used to enhance the quality of mammography images. It also explains the feature extraction and feature engineering process, which involves extracting meaningful information from the images. Furthermore, the chapter defines the Machine Learning models implemented in the research work and provides an overview of the evaluation metrics used.

Towards the end of this chapter, several works related to the theme of breast cancer risk prediction with mammography feature extraction are presented. These studies shed light on existing research and contribute to the overall understanding of the topic.

Chapter 3

Methodology

3.1 Overview

The methodology employed for breast cancer risk prediction involves a series of steps. The process begins with inputting mammographies from the San José hospital as the primary dataset. Comprehensive data filtering techniques are applied to eliminate irrelevant or noisy samples to ensure data quality. Subsequently, the images undergo preprocessing, including normalization, resizing, and mirroring, enhancing comparability and reducing variations. Image segmentation techniques are then employed to identify and isolate regions of interest (ROIs). The next step involves extracting relevant features from the ROIs, utilizing texture analysis and shape descriptors. Feature engineering techniques are applied to enhance the predictive power of the analysis. These include disconnecting side and view information, adjusting features, and combining features from different ROIs.

Moreover, non-radiomic features are incorporated to provide a comprehensive analysis. The dataset is further examined to identify and discriminate duplicated cases. Univariate analysis is conducted using the area under the curve (AUC) metric to evaluate the performance of individual features. Finally, machine learning prediction uses two Cox models, BSWiMS and LASSO. The outcome of this methodology is the prediction of breast cancer risk and related biomarkers, offering valuable insights for clinical decision-making.

3.2 Dataset

This research project is conducted with the Breast Cancer Center at San José Hospital, which supplied the dataset in this study's methodology. The dataset consists of screening mammograms followed up from 2014 to 2016, additional details regarding the image acquisition process, such as the machine used, the force and Kv applied by the mammography machine, and critical patient information, including whether they were diagnosed with cancer during the data collection period. Image 3.2 presents a comparative example of screening images and diagnostic images of a patient.

Initially, the dataset encompassed screening mammograms of 24,037 patients. However, a selection process based on inclusion criteria was applied (depicted in detail in section 3.4.1). The inclusion criteria consisted of several steps: firstly, examining whether patients

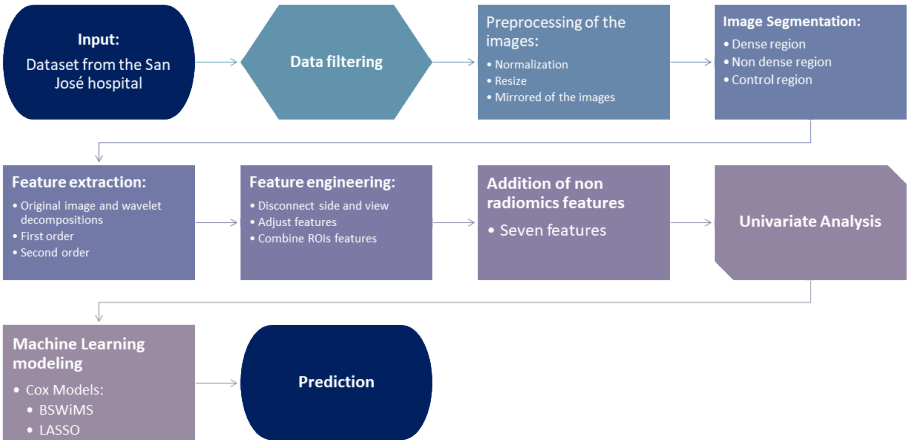

Figure 3.1: Flow of the systematic methodology followed in this research work.

had evidence of undergoing cancer screening for a minimum of four years, regardless of their cancer status; secondly, verifying that the patients were above 40 years of age; thirdly, ensuring that patients did not have breast implants; fourthly, reviewing the radiologist reports to confirm the presence of a BIRADS score of less than 4; and finally, excluding patients who received a cancer diagnosis between 2014 and 2016.

These steps ensured the dataset's reliability, consistency, and relevance to our research objectives. By adhering to these criteria, we aimed to enhance the dataset's quality, enabling accurate analysis and contributing valuable insights to the breast cancer research and detection field.

3.3 Metrics

For this research work the evaluation of the final model will be done with the following metrics:

- Accuracy: It measures the overall correctness of a classification model [94]. It is the ratio of the model's correct predictions to the total number of predictions. Equation 2.4 presents the calculation of accuracy.
- Sensitivity: It is crucial in models where failing to detect a positive instance (a false negative) could have significant consequences [62]. Also known as recall or true positive rate, it quantifies the model's ability to correctly identify positive instances from the overall actual positives. In other words, it measures the proportion of positives correctly identified as such. Please refer to equation 2.5.
- Specificity: Specificity is crucial in models where incorrectly identifying a negative instance as positive (a false positive) can have significant consequences [62]. It quantifies the model's ability to correctly identify negative instances out of all negatives. In simpler terms, it measures the proportion of actual negatives that are accurately identified as such. To calculate specificity refer to equation 2.6.

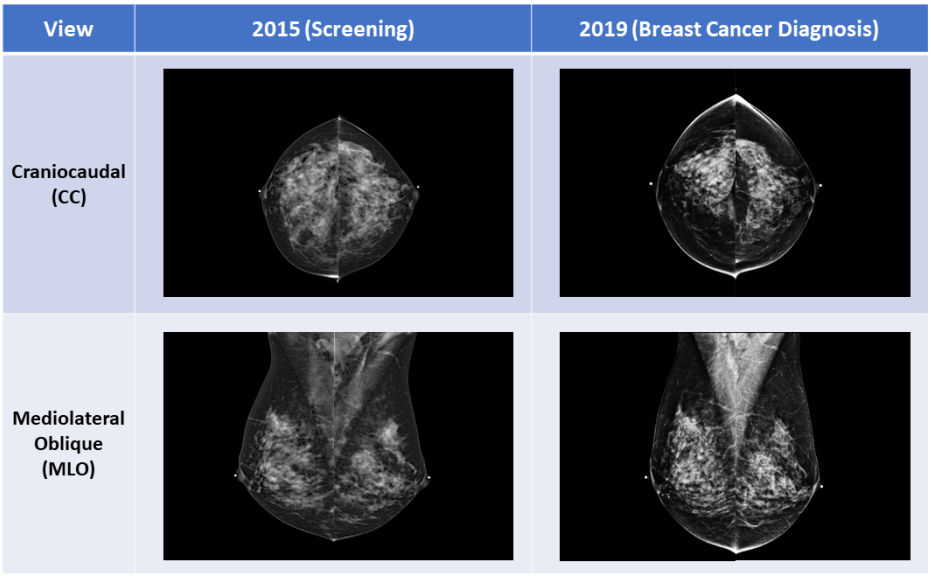

Figure 3.2: Comparative example of screening against breast cancer diagnosis images

- Area Under the Curve: It refers to the area under the Receiver Operating Characteristic (ROC) curve [114]. The ROC curve is a plot that illustrates the true positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold settings. The AUC measures the entire two-dimensional area underneath the ROC curve from (0,0) to (1,1). It provides an aggregate measure of performance across all possible classification thresholds. The AUC metric is beneficial for binary classification problems and evaluating models on imbalanced datasets. An AUC of 1 indicates a perfect model, while an AUC of 0.5 represents a model that performs no better than random chance. An AUC of less than 0.5 suggests a model performing worse than random chance.

Example of the calculation of the metrics

In the following section, an illustration will be provided to demonstrate the computation of the aforementioned metrics, facilitating a better understanding of their application.
The foundation for the metric calculations will be established using the subsequent confusion matrix 3.3:

The first metric we will compute is accuracy, the calculation of which is demonstrated in Equation 3.1. Following that, we will determine Sensitivity, as shown in Equation 3.2. Subsequently, we will calculate Specificity, as detailed in Equation 3.3. We can use these metrics to plot the ROC curve and measure the AUC. These stepwise calculations will allow us to evaluate the model's performance comprehensively.

$$Accuracy = \frac{44 + 211}{44 + 40 + 12 + 211} = \frac{255}{307} = 0.83. \tag{3.1}$$

$$Sensitivity = \frac{44}{12 + 44} = \frac{44}{56} = 0.78. \tag{3.2}$$

$$Specificity = \frac{211}{211 + 40} = \frac{211}{251} = 0.84. \tag{3.3}$$

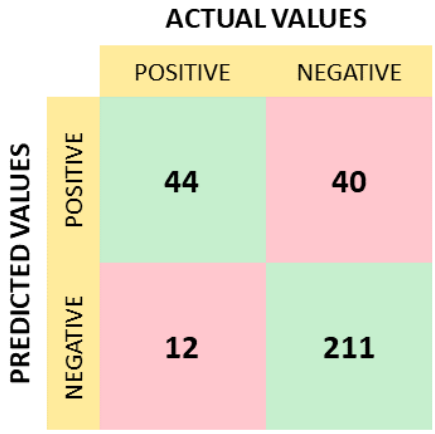

Figure 3.3: Example of a confusion matrix of a model

3.4 Steps of the Methodology

The first step of our research model is to define the input, a carefully selected database obtained from ”El Centro de Cáncer de Mama del Hospital San José.” This database consists of digital mammograms with Medio Lateral-Oblique (MLO) and Cranial-Caudal (CC) views and annotations from specialized radiologists. These mammograms provide detailed images of the breast tissue, allowing us to analyze potential abnormalities and signs of breast cancer. The inclusion of annotations enhances the accuracy and reliability of our research by providing expert guidance and highlighting areas of interest within the mammograms.

3.4.1 Data Filtering

In the second phase of our study, a thorough data filtering process was implemented to refine the dataset provided by Hospital San José. The initial dataset consisted of 24037 cases. Firstly, 697 breast cancer diagnoses were made within 120 days of the initial screening. The remaining 23,340 cases were divided into those with a positive breast cancer diagnosis and those without (control cases). Each group was then subject to specific exclusion criteria; review figure 3.4 as the reference for this analysis.

In the control group of this study, numerous patients did not present proof of remaining cancer-free in the subsequent four years. Several were excluded for various reasons: they were under 40, had breast implants, only underwent preliminary screenings, or had a BIRADS assessment score exceeding 4. The group with a positive diagnosis had the same exclusion criteria. However, another criterion was applied to this group: patients diagnosed in or after 2017 were excluded.

After applying these exclusion criteria, it was applied a matched-case-control study due to the dataset’s imbalance, it was selected a 1:3 scale indicating three negative cases for each positive case. The three control cases were chosen based on similarity to the positive cases, considering factors such as age, BI-RADS score, mammography machine used for image acquisition, and density score.

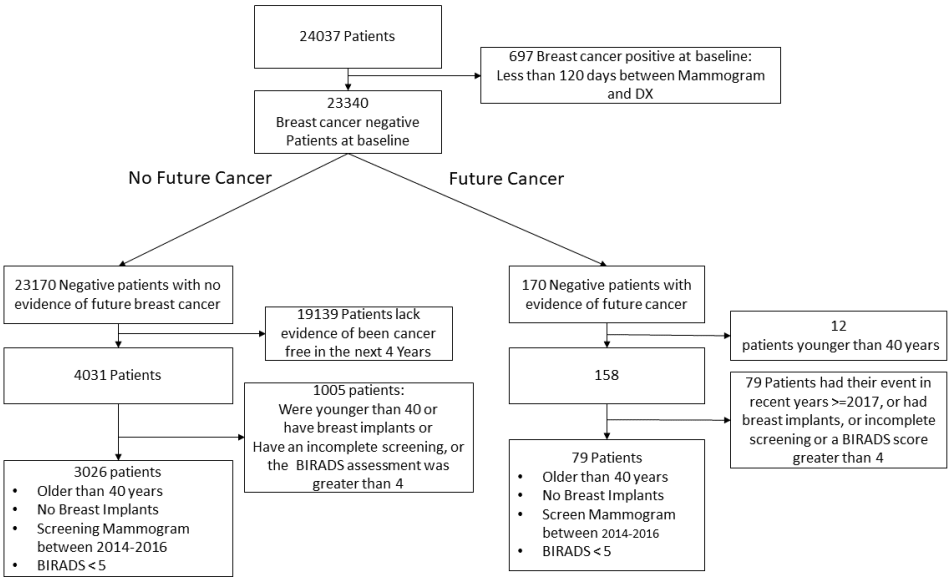

Figure 3.4: Exculsion criteria applied to the dataset of San José hospital

After the matched-case-control study, another exclusion step was conducted, eliminating additional cases with implants and considering only the first screening event for each patient. The dataset's target was the number of days to a cancerous event, calculated using the patient's last mammography for negative cases.

3.4.2 Image Preprocessing

Following the data filtering process, the subsequent step entailed image preprocessing. The preprocessing involved several procedures to ensure uniformity and optimal analysis.

Firstly, pixel normalization was performed to standardize the pixel values across all images. This normalization process allows for accurate and consistent comparisons between different images.

Additionally, the images were resized to ensure uniform size, aligning with the resolution of the original image acquisition. By resizing the images, we achieved consistent dimensions, facilitating efficient processing and analysis.

Furthermore, a mirroring technique was applied for images showing the left side of the breast. The mirrored involved flipping the left side images horizontally to align them with the orientation of the right side images. This step was crucial to ensure that both sides of the breast were equally oriented and enable accurate comparisons and assessments.

Likewise, the image underwent wavelet decomposition using the Haar wavelet, which involved three distinct decomposition levels. From each level, the magnitude was extracted. Consequently, each image now comprises the original image and includes three transformed images resulting from this process. Haar wavelets benefit mammography decomposition due to their simplicity, efficiency, ability to offer multiresolution analysis, effectiveness at edge detection, and noise reduction capabilities.

We optimized the dataset for further analysis and modeling by implementing these

preprocessing techniques. Uniform pixel normalization, image resizing, and mirroring ensured consistent data representation, enhancing the accuracy and reliability of our research outcomes.

3.4.3 Image Segmentation

After completing the preprocessing step, the subsequent stage is image segmentation, where the images are divided into three distinct regions of interest: the control region, the non-dense region, and the dense region.

Our study has delineated a specific zone in the breast known as the control region. This region, a 5-millimeter-wide area positioned in the superficial or peripheral zone of the breast, was identified using our unique methodology. It was assumed that this width was adequate for our research. However, it is essential to note that the probability of detecting cancer within this defined control region is notably low. The reduced likelihood of cancerous development is primarily due to the region's composition, predominantly of fatty tissue, which tends not to be a common site for cancer manifestation.

The breast parenchyma, primarily composed of lobules, ducts, and supporting stroma, is the glandular tissue that constitutes most of the breast. This central structure is encircled by adipose tissue, serving as a supportive and structuring element within the overall anatomy of the breast. The non-dense region within the parenchyma encapsulates this fatty tissue, contributing to the breast's structure and volume. Conversely, the dense region of the parenchyma, which represents the mammary tissue, is attractive to medical professionals as it is where cancerous lesions are more likely to be found. These distinct zones, both non-dense and dense, play different yet crucial roles in the breast's functioning, structure, and health assessment.

The initial segmentation involves differentiating the breast from the background by applying a manual threshold computed from the image histogram. Subsequently, the control region is extracted using morphological operations to segment a 5-millimeter-wide region from the breast. The remaining tissue, the breast minus the control region, is further segmented using a median threshold. This results in segmenting the non-dense and dense regions within the breast.

These segmentation steps yield three distinct and well-defined zones, which can be compared based on visual characteristics. This segmentation process is crucial for our research as it enables us to study and analyze each region separately, providing valuable insights into the different areas of the breast.

3.4.4 Feature Extraction

The feature extraction process can be split into two primary stages: initially, features are extracted from the original image, and subsequently, from three wavelet decompositions of the image. The work of Tamez et al. [137] served as the motivation for utilizing this feature extraction methodology.

In the first wavelet decomposition, we perform operations at level 2 using the Haar wavelet. The second decomposition is computed from the magnitude of the preceding decomposition, again at level 2 with the Haar wavelet. The third and final decomposition follows the same process, using the resultant magnitude from the second decomposition at level 2 with the Haar wavelet. Both stages of the process extract the same types of features, which can be

categorized into two groups: first-order features and second-order features. For more detailed information, refer to Table tables 4.1 and 4.2.

The initial group, first-order features, includes statistical attributes such as the mean, standard deviation, skewness, kurtosis, compactness, among others. These features are itemized in the Table 4.1. The subsequent group, the second-order features, comprises elements from four Gray Level Co-occurrence matrices and features derived from two sizes of the local binary pattern. The specifics of these features are outlined in the second part of Table 4.2.

3.4.5 Feature Engineering

After acquiring the feature matrices, the following crucial step was to conduct feature engineering techniques. This process was necessary to adjust and optimize the features, reducing their number significantly to make future analysis more manageable and less resource-intensive.

Given the vast number of features derived from wavelet decompositions and the Local Binary Pattern, our dataset became hyperdimensional, meaning it contained more features than cases. This complexity necessitated the use of feature-reduction techniques. Such techniques are critical in managing high-dimensional datasets as they help eliminate redundancy and collinearity while preserving essential information. By reducing the data's dimensionality, we can make the dataset more manageable and significantly enhance the efficiency of subsequent machine learning algorithms. These streamlined datasets provide a more viable platform for analysis, improving both the speed and accuracy of our machine-learning tasks.

Furthermore, an innovative approach was adopted during this step, which involved integrating different perspectives of mammography. Specifically, the side and view of the mammography were combined to form a more holistic representation of the mammogram, providing a more comprehensive understanding of the structure and abnormalities, if any.

Additionally, the regions of interest within the mammogram were also combined. This technique can enhance the detection of potential abnormalities by enabling a more comprehensive view of mammographic features. This selection will be made in accordance with the methodology performed by Celaya-Padilla [24].

3.4.6 Addition of Non-Radiomic Features

Upon thorough examination, it was suggested that including features not directly extracted from the images, particularly radiomic features, could significantly bolster the model's performance. By introducing non-radiomic features, the aim was to enrich the dataset further, providing a more comprehensive representation of the information contained within the images. This augmented data complexity is expected to enhance the model's predictive capacity, potentially leading to more precise and accurate results. This approach recognizes the value of diverse data sources in capturing a broader array of relevant factors, thereby strengthening the robustness and generalizability of the model.

Breast cancer risk assessment models can be significantly improved by integrating features like breast density and other factors from mammography acquisition, such as thickness, kV, mA, and force [152]. These features, particularly breast density, play a key role in assessing cancer risk, especially in women with high breast density who might need more intense monitoring or additional screenings [19].

The model becomes adaptable to various mammography systems by including these features, enhancing its usability in different clinical settings[19]. This incorporation not only refines risk assessment and diagnosis accuracy but also improves the selection of parameters and standardization of interpretation. As a result, these comprehensive, personalized risk assessment models can lead to better breast cancer diagnoses and improved patient outcomes.

3.4.7 Univariate Analysis

One important step in the methodology is the univariate analysis of the dataset of previously processed features; as mentioned before, the dataset includes radiomic features computed from the mammograms and the non-radiomic features from the event information. This dataset went through a univariate analysis to examine the characteristics of the dataset of features and review the most significant features. It was selected the AUC as the ranking test for the features. In conclusion, the univariate analysis was a critical tool to understand the data and visualize the most significant features.

3.4.8 Machine Learning Modeling

In the machine learning models we employed, we utilized 369 features per case. These were composed of two distinct categories: 362 radiomics features and seven non-radiomic features. Given that our machine learning models were based on Cox proportional hazards models, they are designed to assess the hazard or risk of an event occurring over time. Therefore, our target variable incorporated the time of an event. For the positive cases, this value is the time elapsed between the screening mammography date and the diagnosis date. For the control cases, it is the time between the screening date and the last recorded screening date. Using Cox models, we can calculate each patient's risk ratio based on the predictor variables' values and the coefficients from the model.

Our study established low and high-risk groups based on a sensitivity threshold derived from each model. This approach categorized patients registered with breast cancer as high risk, while the rest were classified as low risk. The determining factor for this threshold was the inflection point on the Receiver Operating Characteristic (ROC) curve, which offers a balance between sensitivity and specificity in our risk prediction. Conventionally, it is stipulated that 90% of the population is considered low risk and the remaining 10% high risk. However, due to the applied matched-case-control study in our population and the relatively low number of positive cases, we utilized a 50% criteria, evenly dividing the population into low and high-risk groups.

In this project phase, two models were employed: the Bootstrap Step-Wise Model Selection (BSWiMS) and the Least Absolute Shrinkage and Selection Operator (LASSO). Both of these models were brought to life using R code.

BSWiMS was chosen because of its distinctive ability to perform subset selection [135], thereby enabling the exploration of feasible Cox models from a vast array of features [136]. This technique is part of the FRESA.CAD package in R is a supervised model-selection method designed to identify a single statistical model capable of predicting a specific user-determined outcome, which in this case, is risk prediction.

Concurrently, we incorporated the LASSO model. This model, belonging to the GLMNET package, exclusively employs L1 regularization. This process reduces the coefficients by a constant (λ) to facilitate feature selection by eliminating those coefficients lower than λ [138]. LASSO can explore multivariate models comprised of a large dataset of features, allowing for subset selection while investigating feasible Cox models from an extensive set of features.

In addition, we performed a leave-one-out cross-validation for training and testing with 1000 repetitions on both models to ensure the robustness and reliability of their performance [47].

3.5 Prediction

The Machine Learning Models under discussion were assessed, specifically BSWiMS and LASSO regression, using the ROC (Receiver Operating Characteristic) curve. The ROC curve explores the trade-off between the model's sensitivity and specificity [43].

In addition to this, an analysis of the prediction was carried out using a relative risk graph. This graph scrutinizes the relationship between the model's sensitivity and the risk ratio while considering specificity.

Another vital aspect of the evaluation was measuring the model's performance in discriminating between high and low-risk subjects. This differentiation was achieved through the use of Kaplan-Meier plots.

Further, decision curve analysis was plotted to evaluate and compare the clinical utility of the predictive models or diagnostic tests under consideration. This method allows for an insightful comparison of the various models and tests, enhancing our understanding of their potential clinical applications.

3.6 Summary

This chapter offers a comprehensive explanation of the methodology employed in this thesis. It begins with an overview of the study's workflow. Following this, there is a detailed description of the dataset provided by the Breast Cancer Center at San José Hospital, which served as the foundation for this research. Subsequently, we delve into the metrics used to evaluate the performance of the proposed methods, supplemented by an example of the calculations involved. The final part of this chapter meticulously delineates each step carried out in this thesis. After laying out the methodology, we will present the results of each experiment, followed by a thorough discussion and interpretation of these findings. This methodical approach ensures a clear understanding of the study's process, outcomes, and implications.

Chapter 4

Results

4.1 Data filtering

After applying the exclusion criteria illustrated in Figure 3.4, the refined dataset consisted of 3026 patients without a future cancer diagnosis and 79 patients who received a cancer diagnosis.

Given that our samples span 2014, 2015, and 2016, we conducted a matched-case-control study at a 1:3 scale. This study occurred in each year's dataset, considering the results of the aforementioned exclusion criteria. For 2014, we had 32 positive samples, resulting in 96 control samples. In 2015, we had 34 positive samples, leading to 102 control samples, and in 2016, we had 30 positive samples, yielding 90 control samples.

Further refining the dataset, the manual detection of implants led to the exclusion of 10 additional control cases. The resulting population included 96 positive cases and 278 control cases.

In the final stage of this section, we discarded samples from the same patient taken at different dates, opting to retain only the first observation of each patient. This process resulted in a final dataset of 66 samples with a future cancer diagnosis and 251 control samples.

For a visual representation of the data filtering results, please refer to Figure 4.1.

4.2 Image Preprocessing

During this particular step, three distinct processes were involved: a mirroring technique, pixel normalization, and image resizing. To ensure uniformity in orientation, the images from the left mammary glands were mirrored to match the orientation of their corresponding right-side images. Figure 4.2 provides an illustrative example of this mirroring process. For pixel normalization, the following equation (Eq. 4.1) was employed to standardize the pixel values across all images:

$$Normalized = \frac{Original\ image - The\ minimum\ value\ of\ the\ pixels}{The\ maximum\ value\ of\ pixels - The\ minimum\ value\ of\ pixels}. \tag{4.1}$$

In figure 4.3 is presented an example of the original image histogram and the comparison after the normalization. Regarding image resizing, the images were adjusted according to Equation

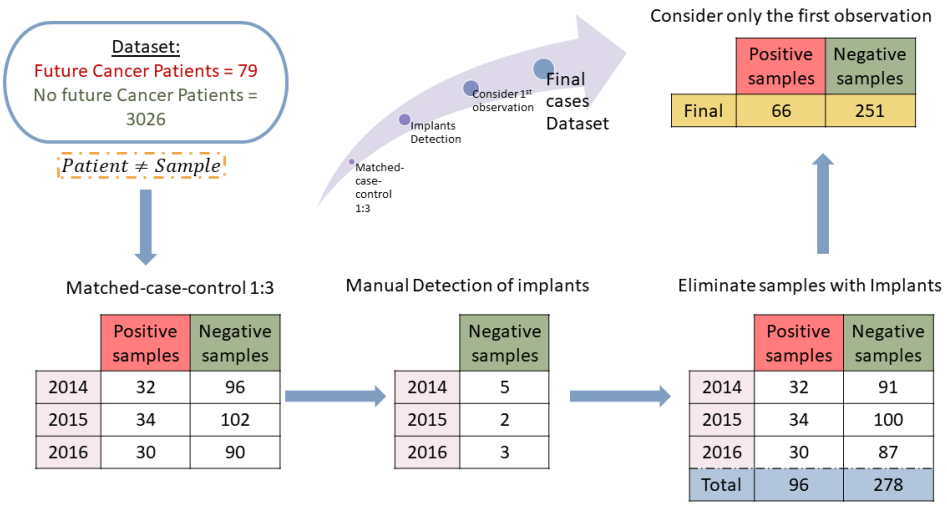

Figure 4.1: Visual representation of the results in the step of Data Filtering.

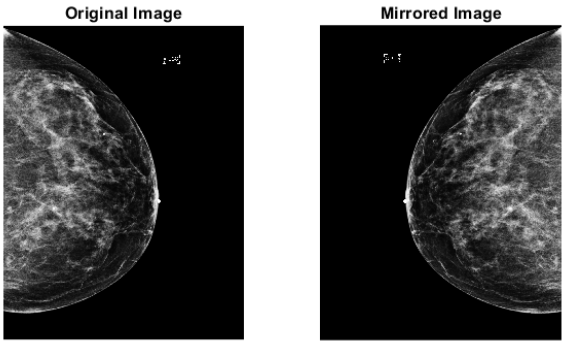

Figure 4.2: Visual example of the mirroring process for left-side mammography.

4.2. Here, f represents the input image, X signifies the pixel spacing (i.e., the physical size of each pixel), and the resolution was set to 0.1 mm.

$$f_{resized} = imresize\left(f, \frac{X}{resolution}\right). \tag{4.2}$$

An illustrative example showcasing the results obtained in this process is presented next.
$Orignal\ Image\ size = 3584 \times 2816\ pixels.$
$Pixel\ Spacing = 0.0814mm.$
$Size\ of\ the\ scaled\ image = 2918 \times 2293\ pixels.$

The final phase in the preprocessing segment involved implementing the wavelet transformation. We decomposed the images using the Haar wavelet and applied them at three distinct levels, generating three additional images. These new images were computed by assessing the magnitude of the wavelet coefficients at each level. An illustration of the wavelet decomposition process is depicted in figure 4.4. This technique is designed to identify specific structures,

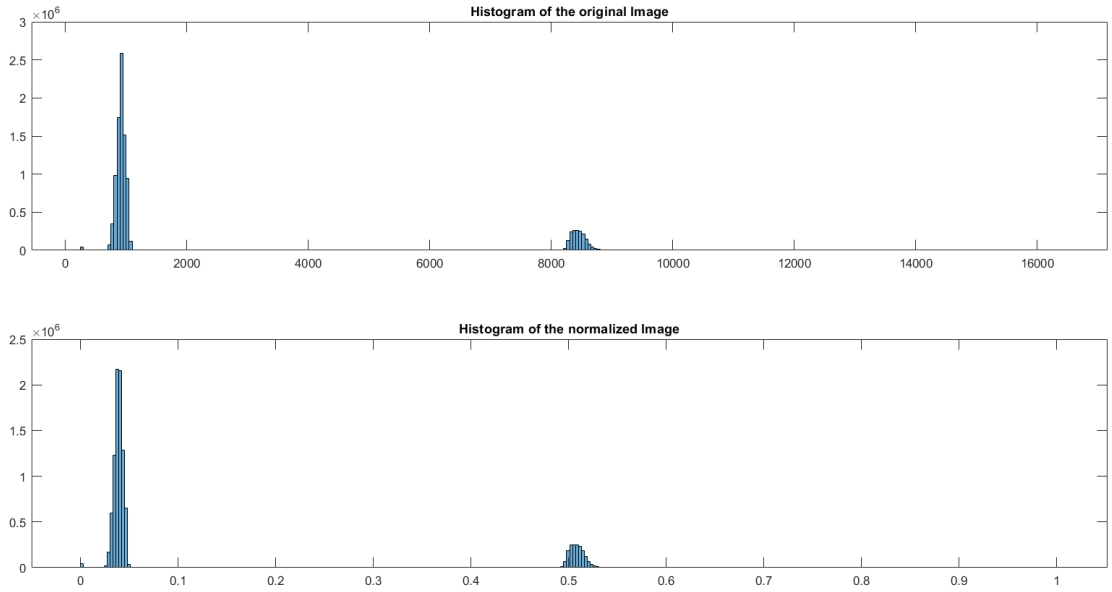

Figure 4.3: Comparison of histograms showing the results of the normalization.

such as masses, with unique frequency characteristics, thereby enhancing the detection and diagnosis of abnormalities in the breast.

4.3 Image Segmentation

The image segmentation using thresholds exhibited exceptional performance, as evidenced in figures figs. 4.5 to 4.7. However, an error is observed in figure 4.8. The control region is visually emphasized in red, the non-dense region is depicted in blue, and the color green represents the dense region in the mammography.

4.4 Feature Extraction

In total, 78 features were extracted, 12 of which are first-order features, and the remaining 66 are second-order features. Feature extraction was accomplished four times for each image, corresponding to the original image and the three wavelet decompositions. Please refer to Tables 4.1 and 4.2 for a detailed overview of the features extracted from each zone of the mammogram.

Each extraction cycle yielded 78 features, leading to 312 features for each region (Dense, Non-dense, Control) within an image. Considering that each mammogram comprises four images, the feature extraction step produced three matrices (one for each zone in the image) with 1248 features for each case.

Features Qty	First Order Features
1	Mean
1	Standard deviation
1	Cube root of the Skewness
1	Fourth root of the Kurtosis
1	Entropy
1	$\log_{10}\left(\frac{\text{std\_dev}\cdot\frac{1}{\lvert\text{mean}\rvert}}{10{,}000}+1\right)$
1	$\log_{10}\left(\frac{\text{quantile in 90}}{\lvert\text{quantile in 50\%}\rvert}\times\frac{1}{10{,}000}+1\right)$
1	$\log_{10}(\text{sum of ROI pixels}+1)$
1	Logarithm of the multiplication of 128*the total sum of the values from the convolution with the 3x3 kernel in the ROI plus one
1	Compactness
1	$\frac{\text{sum of values from convolution with 3x3 kernel}}{\text{sum of values of pixels in ROI}}$
1	$\frac{\text{sum of values from first convolution}}{\text{sum of values from second convolution}}$

Table 4.1: Dictionary of First Order Features extracted

Features Qty	Second Order Features
4	Gray-Level Co-Occurrence Matrix 1: Contrast, Correlation, Energy, Homogeneity
4	Gray-Level Co-Occurrence Matrix 2: Contrast, Correlation, Energy, Homogeneity
4	Gray-Level Co-Occurrence Matrix 3: Contrast, Correlation, Energy, Homogeneity
4	Gray-Level Co-Occurrence Matrix 4: Contrast, Correlation, Energy, Homogeneity
1	Mean of the Contrast of the four GLCM
1	Mean of the Correlation of the four GLCM
1	Mean of the Energy of the four GLCM
1	Mean of the Homogeneity of the four GLCM
23	Local Binary Pattern with diameter 1
23	Local Binary Pattern with diameter 3

Table 4.2: Dictionary of Second Order Features extracted

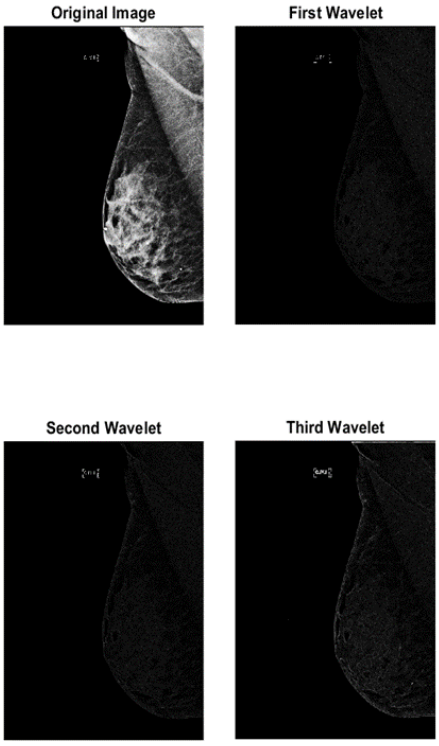

Figure 4.4: Wavelet decomposition Results.

4.5 Feature Engineering

Despite the substantial number of features initially extracted, only two primary feature matrices were considered at this point: the non-dense region and the dense region. Each of these matrices comprised 1248 features across 317 cases, creating a comprehensive feature dataset.

To manage the substantial number of features, we applied feature engineering techniques to streamline and optimize the dataset. This process involved four key steps:

1. Maximizing the three sets of wavelets to capture the most significant features.
2. Reducing the number of features related to the Local Binary Pattern, a texture descriptor.
3. Combining features from both the left and right mammary glands as well as the two viewing perspectives, aligning with the work of Celaya-Padilla et al [24].
4. Combining features from the two segmented regions, specifically the dense and non-dense regions.

Following these steps, the refined feature dataset consisted of 362 features across 317 cases, resulting in a more manageable and focused set of data for subsequent analysis.

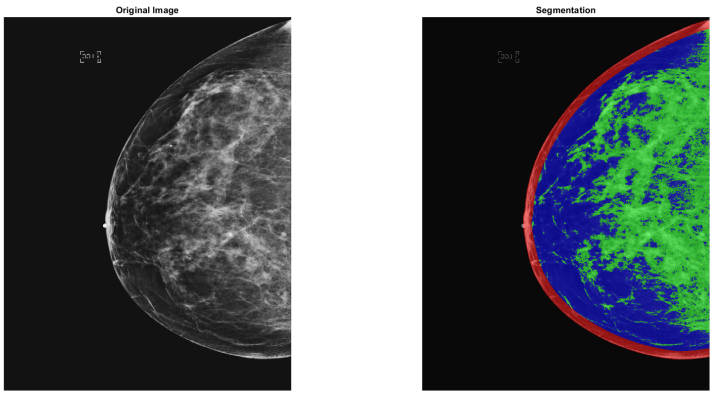

Figure 4.5: Graphical results of the segmentation implemented.

4.6 Addition of Non-Radiomic Features

Accordingly, seven non-radiomic features were integrated into the total feature set. These additional features were derived from the event information and include:

- **Days to Event:** The number of days until the patient was diagnosed with cancer.
- **Age:** The patient's age at the time of the mammogram.
- **Kv:** The kilovolts used in the mammography machine during imaging.
- **Exposure:** The radiation dose the patient was exposed to during the procedure.
- **mA:** The milliamperes used by the mammography machine during imaging.
- **Thickness:** The measurement represented the thickness or depth of the breast tissue being imaged.
- **Force:** The pressure applied to the breast during the mammography procedure.

Following the inclusion of non-radiomic features, the dataset's composition was further expanded. It now encompassed a total of 369 features spread across 317 cases. This addition has further enriched the data, providing a more comprehensive set for analysis.

4.7 Univariate Analysis

The Univariate Analysis of the features involved the utilization of the AUC metric for evaluating the discriminatory power of each feature. Through this analysis, Table 4.3 was generated, presenting a comprehensive ranking of the most discriminative features based on their respective AUC values.

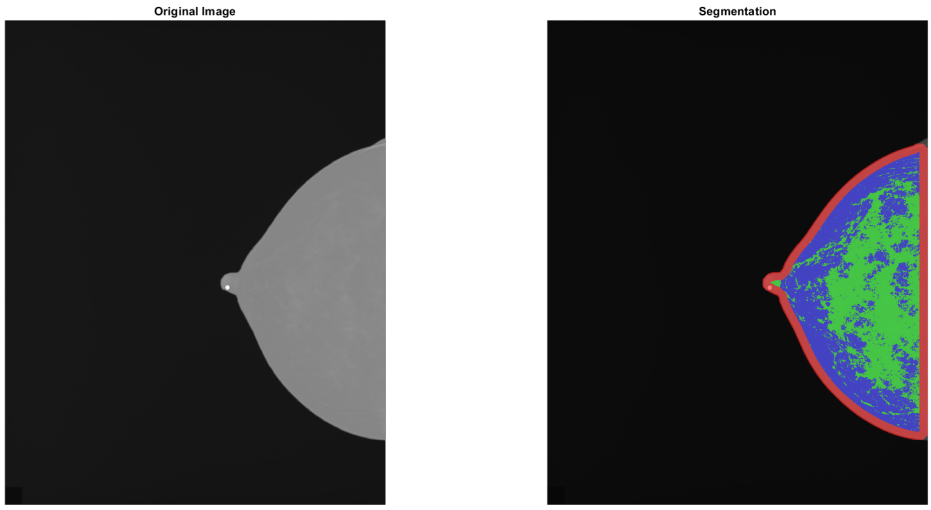

Figure 4.6: Graphical segmentation results implemented with a different vendor utilized for the image acquisition.

4.8 Machine Learning Modeling

The results section of this step will be divided into three distinct parts. Firstly, the outcomes obtained from the BSWiMs model will be presented, highlighting its specific findings and implications. Following that, the results from the LASSO model will be detailed, providing insights into its performance and outcomes. Lastly, the Leave-one-out Cross-validation performance evaluation will be discussed, incorporating the BSWiMs and LASSO models.

4.8.1 BSWiMS

Three graphics were utilized to present the BSWiMS model results: the ROC curve 4.9, the Relative Risk plot 4.10, and the Kaplan-Meier curve 4.11. These graphical representations provide a comprehensive visual depiction of the model's performance, the relative risk between high-risk and low-risk groups, and the survival probabilities of the two groups of cases.

The Relative Risk graphic illustrates the calculation of Risk Ratios for various sensitivity values. The blue vertical line represents the automatically computed threshold, corresponding to the inflection point of the previously presented ROC curve. This threshold holds significance as it helps determine the point at which the Risk Ratio shifts, indicating a change in the relative risk between high-risk and low-risk groups.

The Kaplan-Meier curves clearly illustrate a significant differentiation between the high-risk and low-risk groups when evaluating the survival probabilities of each group. In addition, these curves provide a visual representation of the distinct survival experiences observed between the two groups, highlighting the varying probabilities of survival over time.

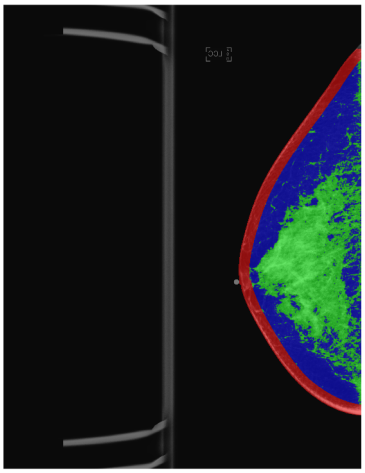

Figure 4.7: Segmentation performance with noisy artifacts

4.8.2 LASSO

To present the results of the LASSO model, three graphics were employed: the ROC curve (Figure 4.12), the Relative Risk plot (Figure 4.13), and the Kaplan-Meier curve (Figure 4.14). These graphical illustrations offer a comprehensive visual representation of the model's performance, the relative risk assessment between high-risk and low-risk groups, and the survival probabilities associated with each patient group.

The Relative Risk graphic displays the computation of Risk Ratios across different sensitivity values. A blue vertical line indicates an automatically calculated threshold, which aligns with the inflection point of the previously shown ROC curve. This threshold is critical as it signifies the point at which the Risk Ratio undergoes a shift, indicating a notable change in the relative risk observed between the high-risk and low-risk groups.

The Kaplan-Meier curves clearly and visually compellingly demonstrate the substantial distinction between the high-risk and low-risk groups when assessing their survival probabilities. In addition, these curves serve as powerful tools for illustrating the marked differences in survival experiences between the two groups, emphasizing the varying probabilities of survival throughout the study.

4.8.3 Leave-One-Out Cross-Validation

In addition to evaluating the two previous models, the results of leave-one-out cross-validation with the BSWiMS model as the predictive model and 1000 repetitions are presented. This cross-validation approach thoroughly assesses the model's performance by systematically leaving out one data point at a time and repeating the process 1000 times. Please refer to Figures 4.15, 4.16, and 4.17 for the corresponding graphics.

Furthermore, a calibration analysis was performed to evaluate the accuracy and reliabil-

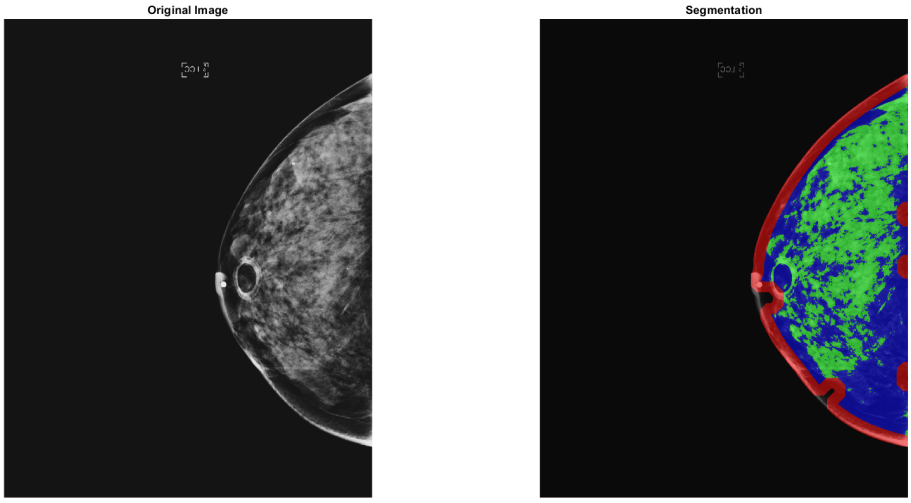

Figure 4.8: Segmentation with not perfect performance

ity of the cross-validation results. Calibration ensures that the predicted probabilities from the BSWiMS model align well with the observed probabilities in the validation dataset. By examining the calibration, we can assess whether the model's predictions are calibrated and reliable, providing insight into the model's overall performance and applicability to new data. The results obtained through this calibration are shown in figures figs. 4.18 and 4.19 and 4.20. After calibration, three key features were identified in the prediction model: DIS_cov (covariance of the non-dense region), Thickness (depth of breast tissue), and DIS_LBP_W1_Wt_sd (Local Binary Pattern Wavelet feature).

Figure 4.21 presents a heat map of these features, with a color bar on the left side indicating blue for control cases and red for cancer events. The heat map provides a visual representation of the distribution and significance of these features in differentiating between control and cancer cases.

	caseMean	caseStd	controlMean	controlStd	controlKSP	ROCAUC	wilcox.Zvalue
ML_DV_glcm1_corr (Mean (R&L) of the subtraction (CC-MLO) of correlation from GLCM1)	2.5458	2.9294	1.7061	3.3585	1.48E-07	**0.618**	2.73
DL_MaDL_glcm1_corr (Subtraction of R&L of the max CC&MLO of correlation from the glcm1)	1.9327	2.1505	1.3194	1.7390	1.81E-12	**0.618**	2.72
ML_DV_glcm3_corr (Mean (R&L) of the subtraction (CC-MLO) of correlation from GLCM)	2.8612	2.9241	2.2252	3.7626	9.32E-07	**0.616**	2.67
ML_DV_glcm2_corr (Mean (R&L) of the subtraction (CC-MLO) of correlation from GLCM2)	2.5032	2.6451	1.8427	3.3027	7.12E-07	**0.615**	2.65
DIS_ene_mean (Mean of the energy from the non-dense zone)	0.0310	0.0280	0.0379	0.0245	2.00E-02	**0.615**	2.64
DIS_cov (Mean of the covariance from the non-dense zone)	0.0943	0.0547	0.1208	0.0770	6.05E-03	**0.613**	2.59
DIS_q90cov (Mean of the covariance from the quantile 90 from the non-dense zone)	0.1139	0.0714	0.1451	0.0986	2.36E-05	**0.613**	2.59
ML_DV_glcm4_corr (Mean (R&L) of the subtraction (CC-MLO) of correlation from GLCM4)	3.4905	3.7520	2.6511	4.2792	8.37E-06	**0.611**	2.53
Thickness	70.9242	12.5224	65.8008	12.5800	3.11E-01	**0.606**	2.41
DL_MaDL_corr_mean (Subtraction of R&L of the max CC&MLO of correlation from all GLCM)	1.9837	1.4726	1.5208	1.2834	1.22E-05	**0.603**	2.33

Table 4.3: Top features obtained from the univariate analysis, ranked by the Area Under the Curve

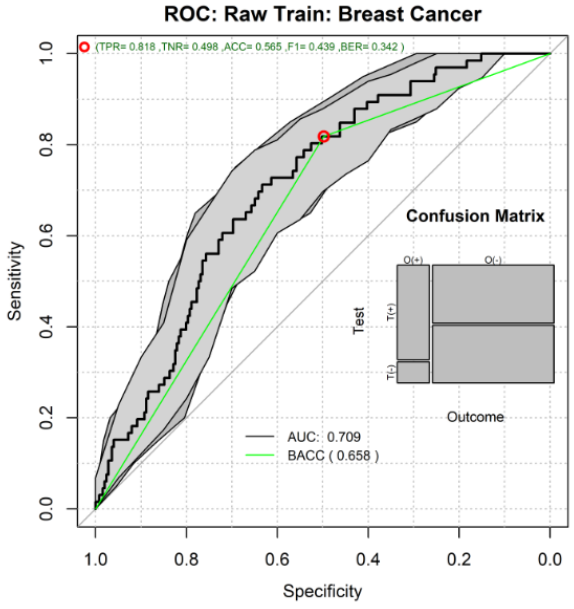

Figure 4.9: ROC plot of the BSWiMS performance.

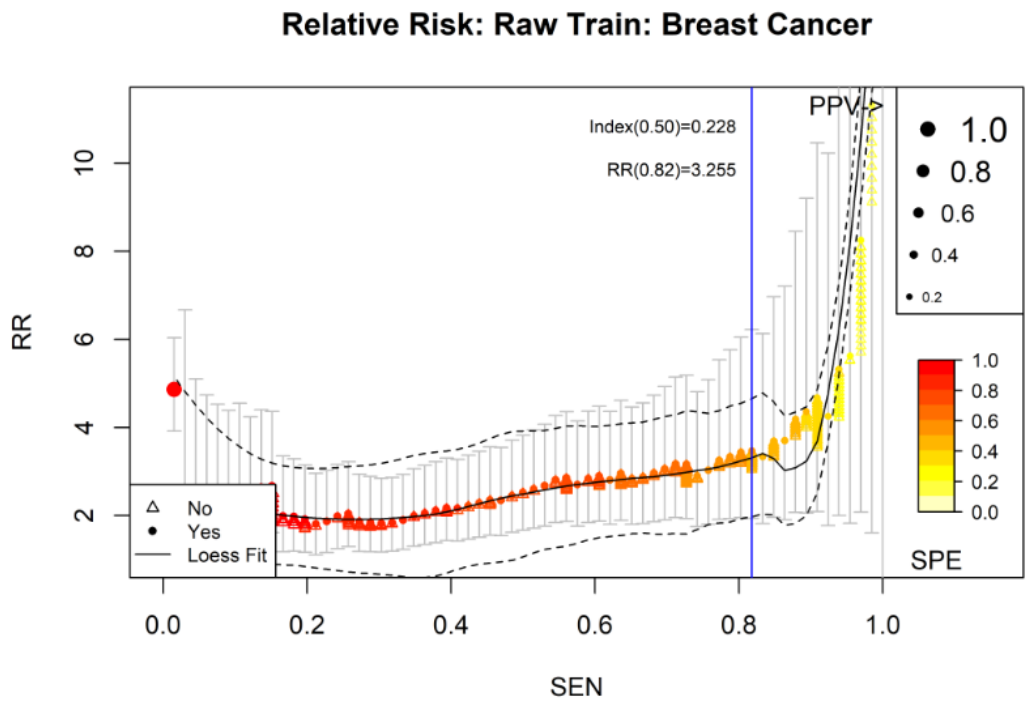

Figure 4.10: The Relative Risk plot of the BSWiMS model.

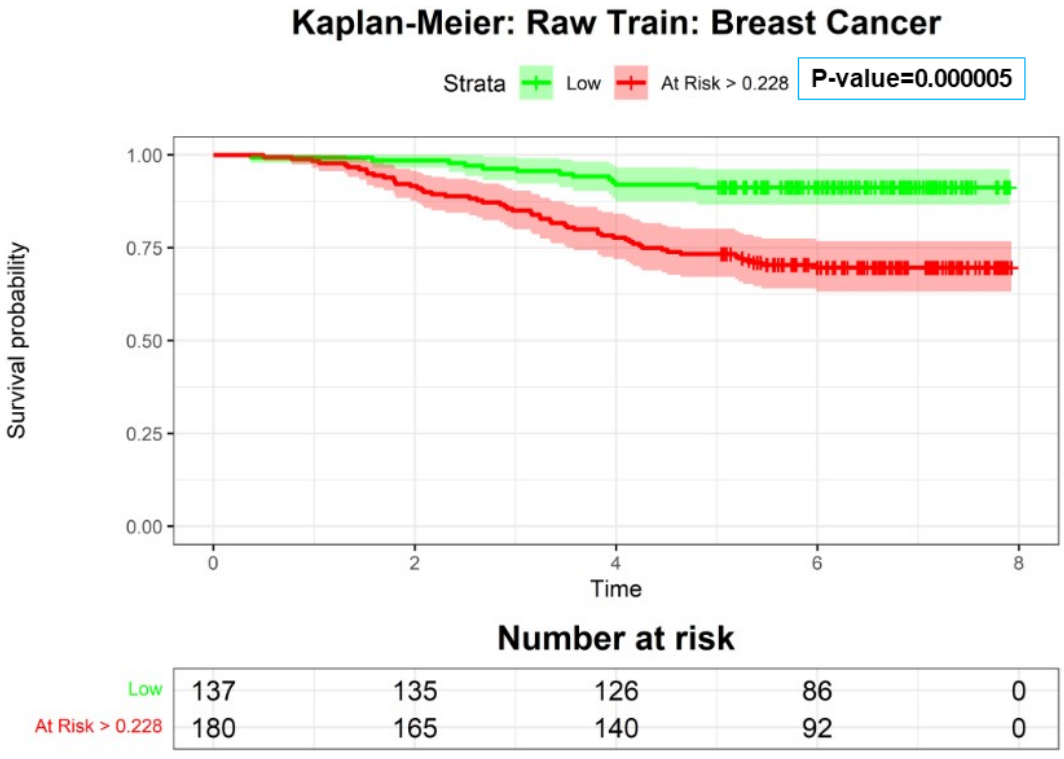

Figure 4.11: The Kaplan-Meier curve of the BSWiMS model.

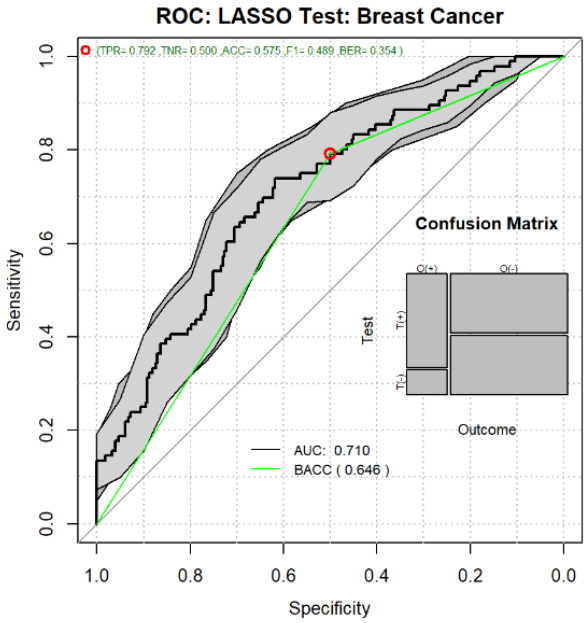

Figure 4.12: ROC plot of the LASSO performance.

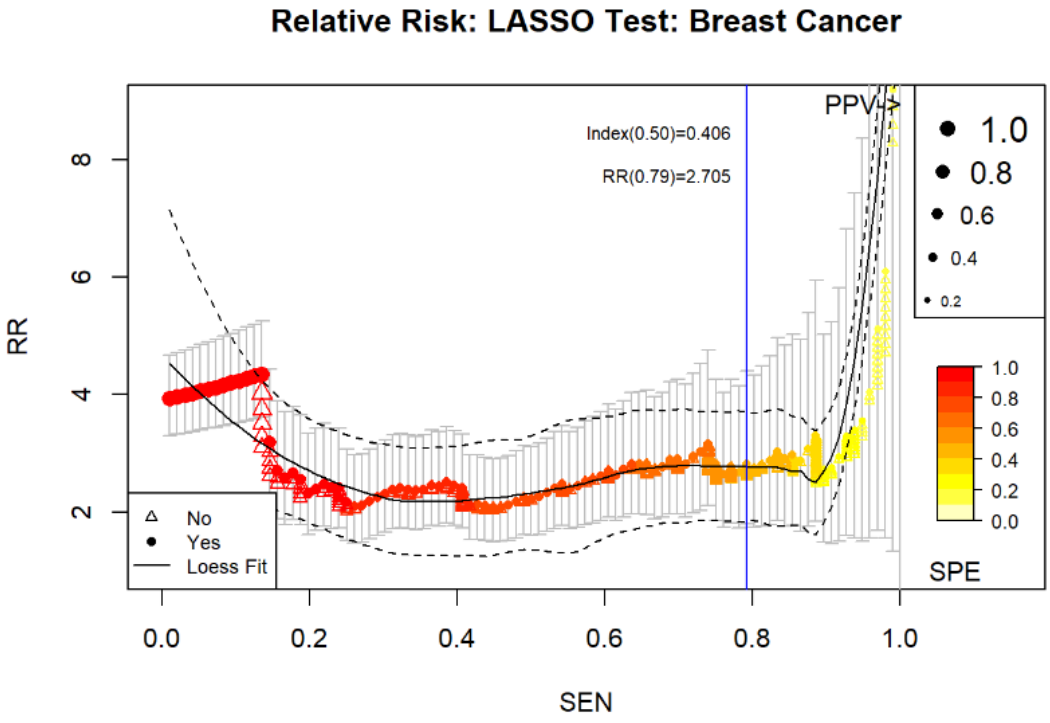

Figure 4.13: The Relative Risk plot of the LASSO model.

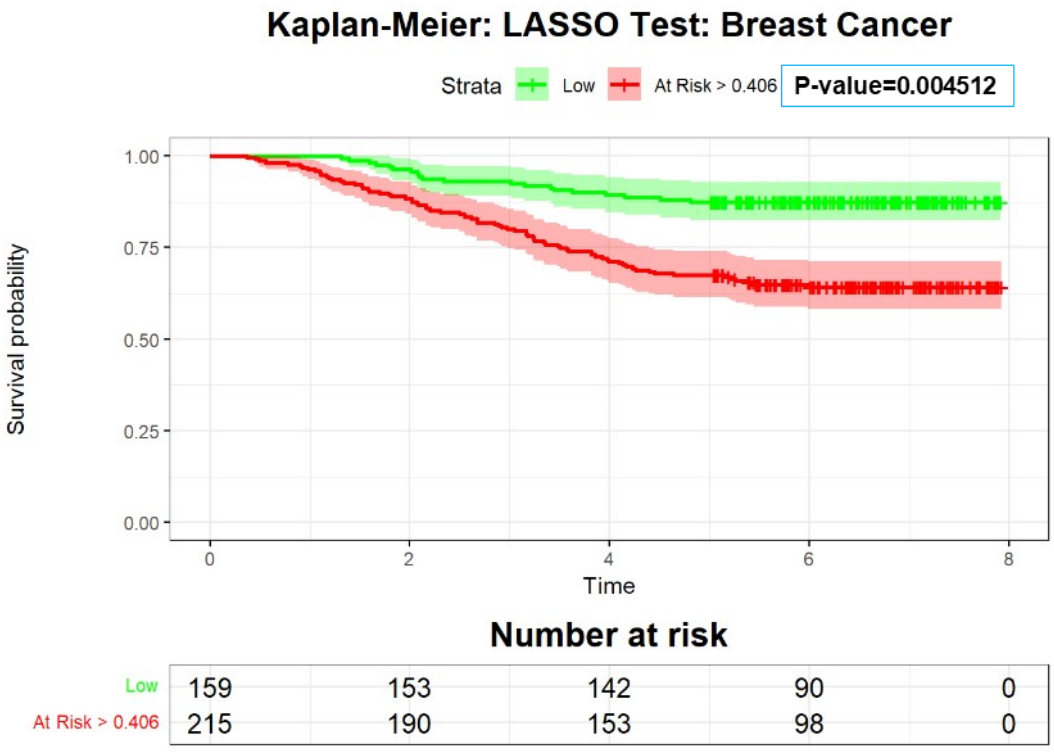

Figure 4.14: The Kaplan-Meier curve of the LASSO model.

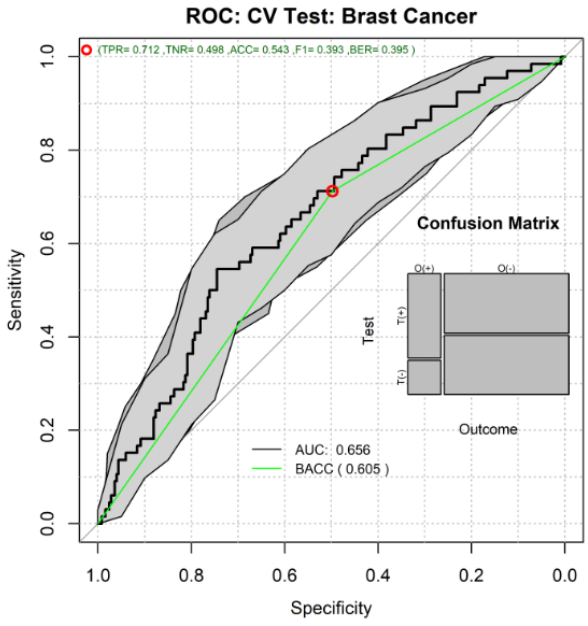

Figure 4.15: ROC plot of the Leave-one-out Cross-Validation with the BSWiMS model performance.

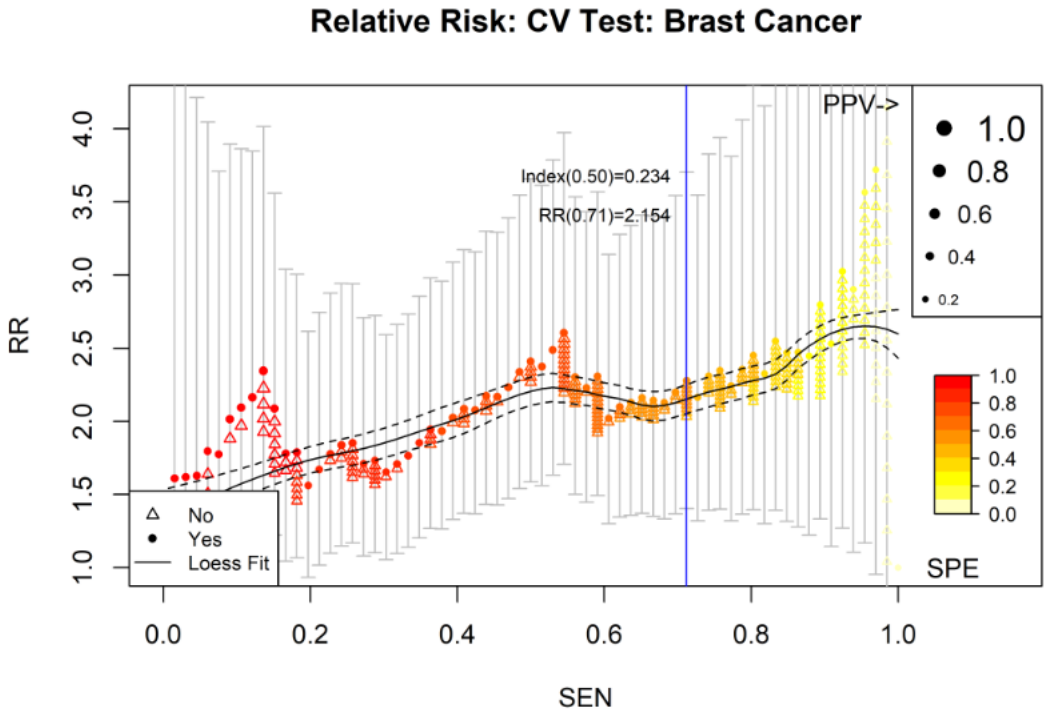

Figure 4.16: The Relative Risk plot of the Leave-one-out Cross-Validation with the BSWiMS model.

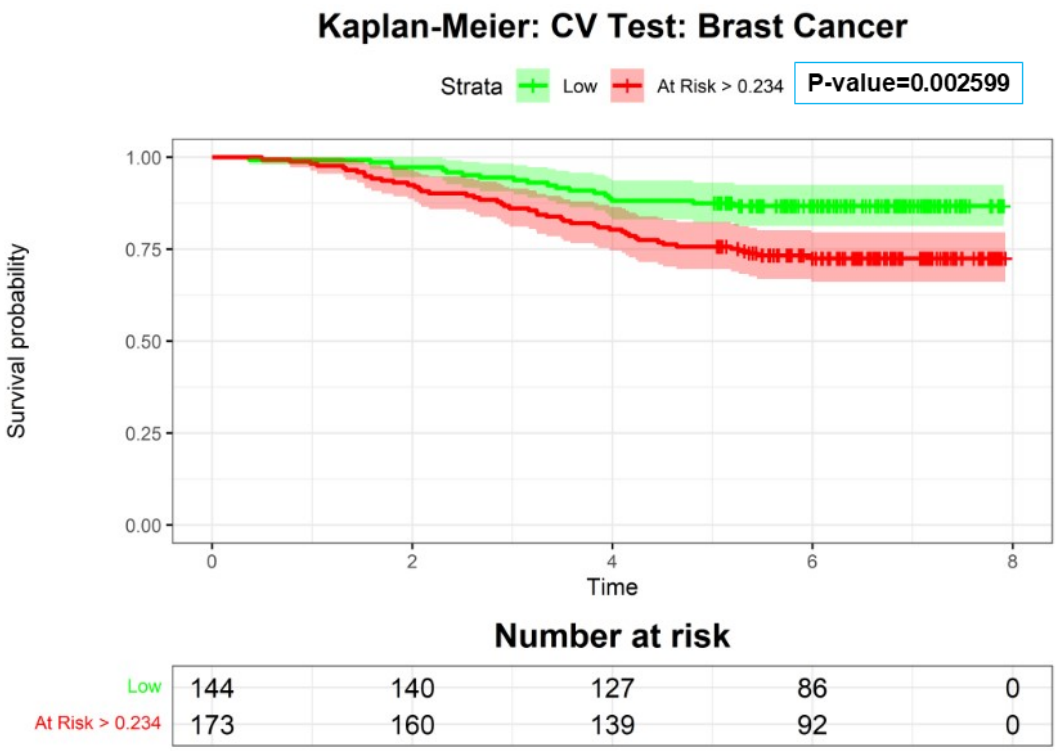

Figure 4.17: The Kaplan-Meier curve of the Leave-one-out Cross-Validation with the BSWiMS model.

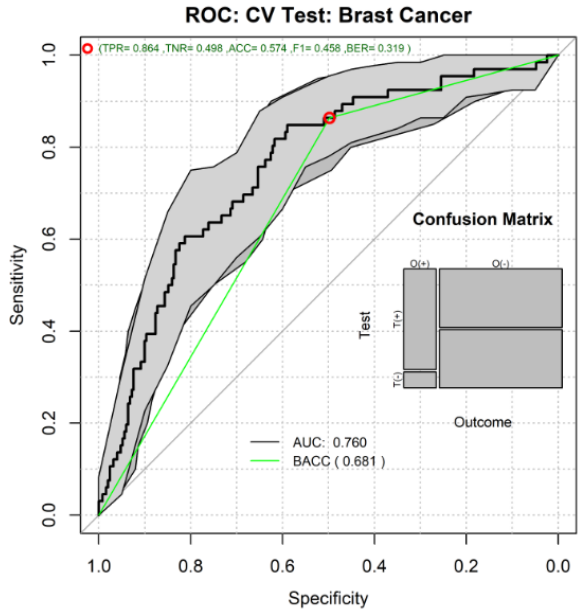

Figure 4.18: ROC plot of the Leave-one-out Cross-Validation with the calibrated BSWiMS model performance.

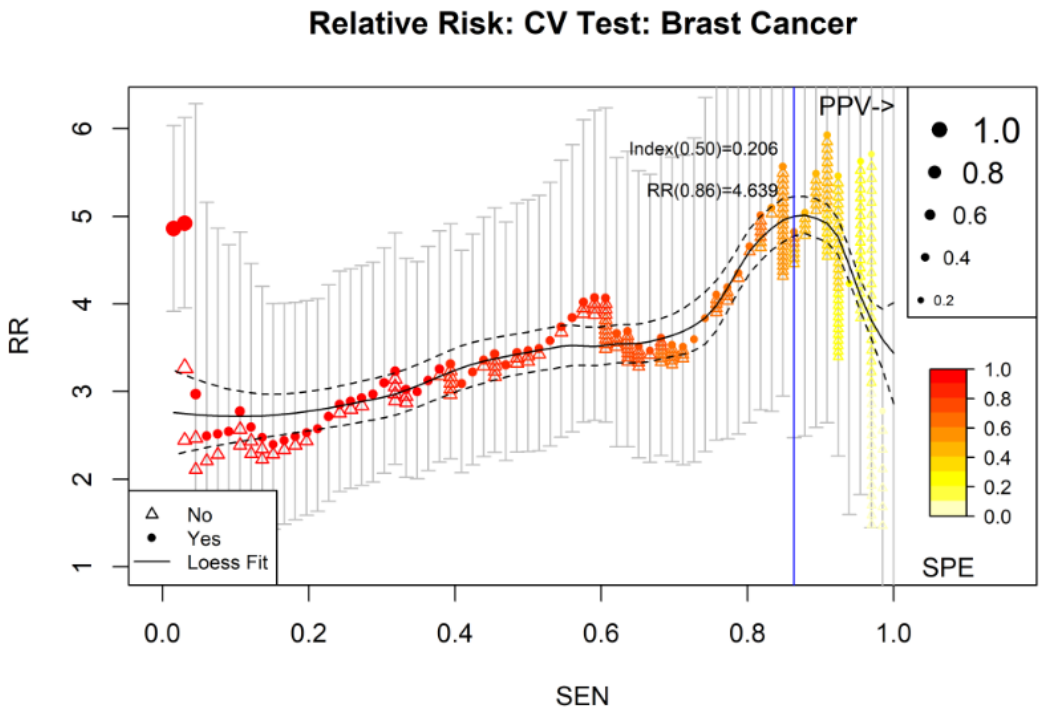

Figure 4.19: The Relative Risk plot of the Leave-one-out Cross-Validation with the calibrated BSWiMS model.

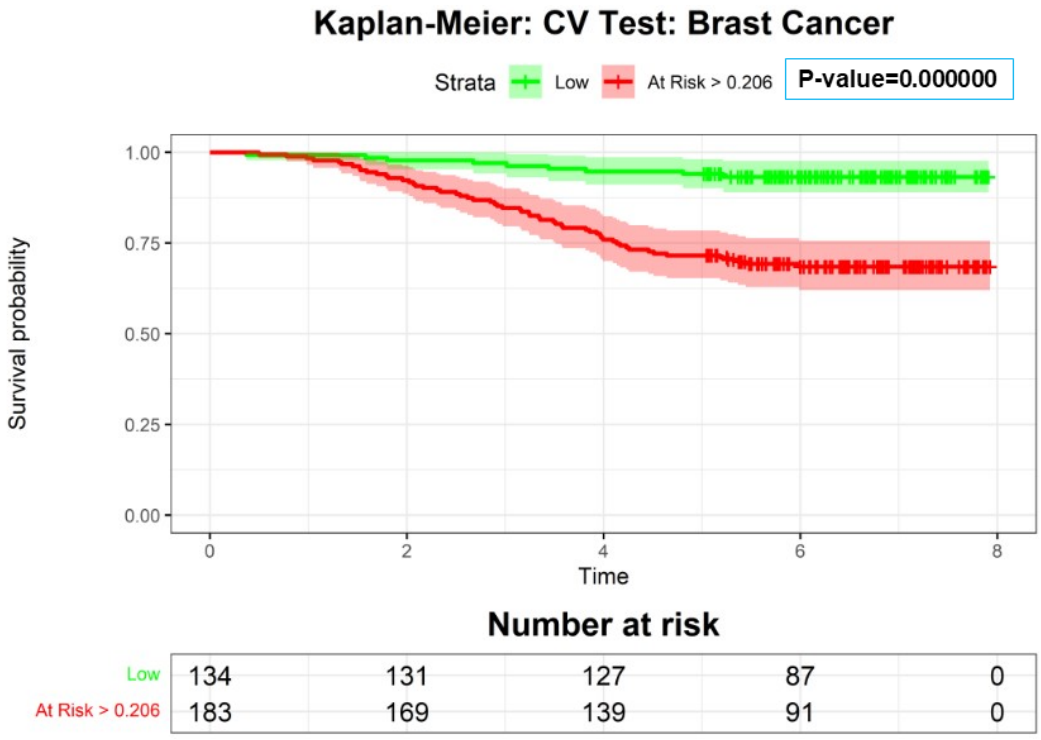

Figure 4.20: The Kaplan-Meier curve of the Leave-one-out Cross-Validation with the calibrated BSWiMS model.

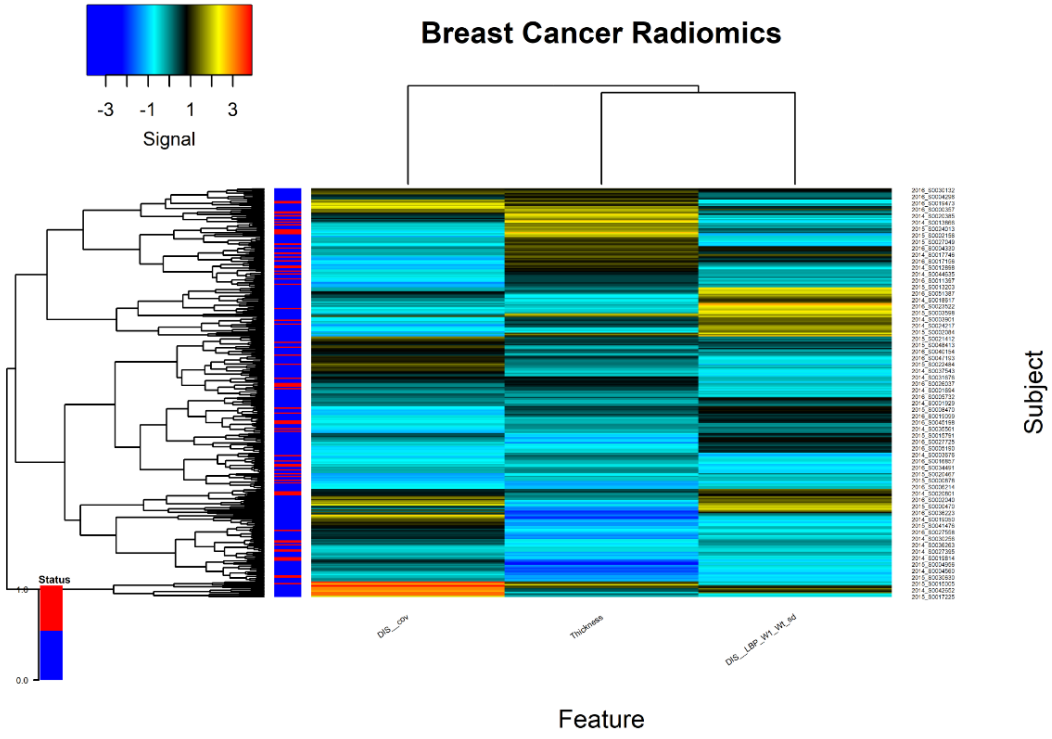

Figure 4.21: Most selected features in the model with Cross-validation.

Chapter 5

Discussion

Processing of Images

Multiple steps were undertaken during the image processing stage, including normalization, segmentation, and wavelet transformation. The normalization process was successfully carried out despite working with a dataset containing mammography images from different vendors with varying pixel value ranges. The segmentation of the images generally yielded satisfactory results, although challenges were encountered with images lacking clear boundaries of the mammary regions. Additionally, cases with implants not initially documented in the mammography reports were identified, and some images exhibited noisy artifacts. These challenges were addressed by manually calculating thresholds specific to the characteristics of images from different vendors and excluding cases with implants. Furthermore, a three-level decomposition using the *'Haar'* wavelet was employed. With each level doubling the number of images compared to the previous level, 64 images were generated. However, for further analysis, only three images were considered by extracting the magnitudes at each level, thereby reducing the number of images.

Feature Extraction

The feature extraction process yielded a total of 78 features per image. Considering that each mammography case consisted of four images (two for each breast at CC and MLO projections), and each image underwent wavelet decomposition resulting in three transformed images per view, the feature extraction was performed on 16 images per case. Consequently, the feature extraction technique produced a matrix comprising 1248 features by 317 cases.

During this step, we encountered a challenge where some extracted features resulted in "Not a Number" (NaN) values. This issue arose due to cases of breast implants not being reported in the dataset. To address this, these specific cases were excluded from further analysis, ensuring the integrity and reliability of the feature matrix. By excluding cases with breast implants, the impact of missing or invalid data was mitigated in the subsequent analysis and interpretation of the results.

Feature Engineering

Feature engineering techniques were employed to reduce the dimensionality of the feature matrix to handle the issue of having more features than cases. The larger group, Local Binary Pattern (LBP) features, and the features obtained from wavelet decomposition were summarized. Since feature extraction was conducted individually on each mammography image, it was crucial to analyze the features in terms of symmetry across the four provided images. The research focus aimed to obtain a comprehensive summary of the patient's features rather than considering the specific side of the breast. This challenge was accomplished by applying the methodology proposed by Celaya-Padila et al. [24], incorporating the bilateral symmetry of the mammographic images. These feature engineering techniques, including summarizing the LBP features and utilizing symmetry analysis, effectively reduced the dimensionality of the feature matrix. As a result, the analysis and interpretation of the results were enhanced, providing a concise and informative summary of the patient's features.

Top Features

The three most discriminative features selected by the model with cross-validation were: DIS_cov (covariance of the non-dense region), Thickness (depth of breast tissue), and DIS_LBP_W1_Wt_sd (Local Binary Pattern Wavelet feature).

The covariance of pixel intensities in an image is a valuable feature that provides insights into the variability and correlation within the image. High covariance values indicate a strong correlation or similarity in pixel intensities, which may indicate the presence of consistent structures or abnormalities. Conversely, low covariance values suggest regions with less correlation or more significant heterogeneity in pixel intensities. In the context of breast imaging, the covariance of the non-dense region, representing the adipose tissue in the breast, can help characterize the image and potentially identify abnormalities associated with a future cancer diagnosis. By analyzing the covariance of mammography images, clinicians can gain valuable information for early detection and diagnosis of breast cancer.

In mammography images, breast thickness represents the depth of the imaged breast tissue. An increased thickness could signify a greater breast density or denser parenchyma, potentially leading to the detection of dense structures often associated with breast abnormalities. Furthermore, this thickness can be influenced by a patient's sensitivity and pain tolerance during the mammography procedure, which involves applying force to compress the breast. Such sensitivity could potentially be indicative of an abnormality or a cancerous lesion. However, the specific mechanisms linking these factors require more in-depth research. Therefore, evaluating breast thickness is crucial for gaining insights into breast density, which holds significant implications for detecting breast cancer and assessing associated risk.

The wavelet decomposition of mammography images provides a multi-scale representation, capturing different levels of details and textures. By applying the local binary pattern (LBP) feature extraction to the wavelet coefficients, the local structural properties can be analyzed at various scales, enhancing the detection of potential abnormalities or markers of breast cancer. This approach enables the identification of specific texture patterns associated with breast cancer, enhancing the accuracy of detection and diagnosis. The combination of wavelet-based decomposition and LBP feature extraction offers a valuable tool for improving

the effectiveness of mammography analysis and aiding in the early detection of breast cancer in clinical settings.

These features capture essential information about the relationship between different regions in the mammographic image, the depth of breast tissue being imaged, and the present textural patterns. Their selection highlights their significance in predicting and assessing breast cancer risk. These findings provide valuable insights for improving screening and diagnostic approaches in the future.

Machine Learning Cox models

Applying machine learning techniques, specifically, Cox models, has demonstrated encouraging results in accurately predicting high and low-risk patients. To validate the reproducibility of these results, a leave-one-out cross-validation was conducted, yielding more impressive results than when training the individual BSWiMS and LASSO models on the raw dataset. An additional calibration step further boosted the prediction models' performance. Notably, these models achieved a significant area under the curve (AUC) of 0.7, exceeding the performance of the MIRAI model [152], which utilized deep learning techniques and had an AUC of 0.6 for this population, as suggested in the literature.

This performance is compelling compared with the findings of Arasu et al. [9], who compared AI algorithms for breast cancer risk prediction, which used mammography, against traditional models like the Breast Cancer Surveillance Consortium. The AI models achieved an AUC ranging from 0.63 to 0.67. Meanwhile, the best traditional risk model, the Tyrer-Cuzick model, has an AUC close to 0.7 in its current version [37].

These observations underscore the effectiveness and reproducibility of the machine learning-based Cox models in correctly identifying patient risk levels. These models present an exciting prospect for improving risk assessment in clinical practice. By providing crucial insights for patient management, they hold the potential to enhance overall healthcare outcomes significantly.

Explainable AI

This study has yielded promising results by implementing radiomics, which enables extracting the most discriminative features. Once these top features are identified, they can be correlated with physical or biological characteristics associated with the risk of developing breast cancer. This comprehensive characterization and scaling of the risk of developing breast cancer can assist physicians in making informed decisions for their patients. For instance, high-risk patients can undergo more frequent screenings to detect breast cancer at early stages.

Study Limitations

Some limitations should be acknowledged in this research work. Firstly, the population of the dataset is small, which calls for additional studies involving more extensive and diverse populations to ensure the generalizability of the findings. Moreover, numerous cases with breast implants pose a limitation, as this study focuses solely on cases without implants. Consequently, the results may not directly apply to individuals with implants, warranting further investigation in that population.

Another limitation stems from the lack of a robust culture of breast cancer screening in Mexico, resulting in patients attending only one screening and discontinuing follow-up. This lack of culture hinders the ability to gather longitudinal data and observe changes in risk over time.

Lastly, the segmentation technique used in this research would benefit from refinement and evaluation by domain experts. Expert input can enhance the accuracy and reliability of the segmentation results, ensuring the proper delineation of regions of interest.

Summary

This chapter provides a comprehensive discussion of the research methodology and results. It covers image processing, feature extraction, feature engineering, top feature analysis, comparative evaluation of machine learning models, explainable AI, and limitations. First, the image processing steps are detailed, followed by a focus on the extracted features and their relevance. Next, feature engineering techniques and their results are analyzed. The discussion emphasizes the most discriminative features and their potential associations with patient characteristics. Next, a comparative analysis using AUC and exploring explainable AI is presented. Lastly, limitations are discussed.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In conclusion, the proposed methodology has been validated, demonstrating promising results in predicting breast cancer risk. In addition, this research provides valuable insights for physicians to make informed decisions regarding the frequency of check-ups, particularly for high-risk patients, aiming to enhance early detection and improve survival rates.

The main contributions of this study lie in the extraction of radiomics features from screening mammographies, and the utilization of machine learning Cox models focused on survival analysis, which effectively identifies the most discriminant features. This implementation serves as a foundation for further investigation in this field and has the potential to impact clinical practice positively.

The implemented models exhibited favorable performance metrics, including area under the curve, sensitivity, and specificity, widely accepted and utilized in the clinical field. These models offer valuable support to healthcare professionals in making informed decisions regarding cancer screening intervals, benefiting high-risk individuals by increasing the likelihood of detecting cancer early and improving patient survival rates.

However, certain refinements are necessary to integrate these models into clinical practice fully. Increased dissemination of such studies is required to foster a breast cancer screening culture within society and encourage greater participation in dataset collection. Furthermore, segmentation, feature extraction, and other analytical aspects should be further refined to enhance model performance.

In summary, this thesis highlights the potential of applying radiomic feature extraction and machine learning Cox models for breast cancer risk prediction, contributing to advancing personalized medicine approaches in breast cancer management.

6.2 Future Work

In terms of future work, several steps can be undertaken to enhance the outcomes of this thesis project further. Firstly, expanding the dataset by incorporating a more extensive and diverse sample would provide a broader representation of breast cancer cases and improve the

generalizability of the findings. Additionally, exploring the extraction of additional image-based features could offer valuable insights into the characterization of various biological and physical aspects of breast anatomy, thereby enhancing the comprehensiveness of the analysis. Furthermore, utilizing the San José dataset, particularly in conjunction with related studies like MIRAI by Yala et al. [152], would enable comparative analyses and facilitate benchmarking to assess the performance and validity of the proposed methodology. In addition, refining the existing steps using advanced techniques such as deep learning could be pursued to improve the accuracy and efficiency of the segmentation process. Finally, testing the developed methodology on a dataset comprising a population with diverse characteristics would comprehensively evaluate its robustness and applicability across different demographics. By undertaking these future steps, this research has the potential to advance breast cancer risk prediction and contribute to the development of more effective and tailored screening approaches.

Bibliography

[1] Abdel Razek, A. A. K., Alksas, A., Shehata, M., AbdelKhalek, A., Abdel Baky, K., El-Baz, A., and Helmy, E. Clinical applications of artificial intelligence and radiomics in neuro-oncology imaging. *Insights into Imaging 12*, 1 (2021), 1–17.

[2] Aerts, H. J. The potential of radiomic-based phenotyping in precision medicine: a review. *JAMA oncology 2*, 12 (2016), 1636–1642.

[3] Aerts, H. J., Velazquez, E. R., Leijenaar, R. T., Parmar, C., Grossmann, P., Carvalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D., et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications 5*, 1 (2014), 4006.

[4] Aggarwal, N., and Agrawal, R. First and second order statistics features for classification of magnetic resonance brain images. *Journal of Signal and Information Processing Vol.3 No.2* (2012).

[5] Aibar, L., Santalla, A., López-Criado, M., González-Pérez, I., Calderón, M., Gallo, J., and Fernández-Parra, J. Clasificación radiológica y manejo de las lesiones mamarias. *Clínica e Investigación en Ginecología y Obstetricia 38*, 4 (2011), 141–149.

[6] Akram, M., Iqbal, M., Daniyal, M., and Khan, A. U. Awareness and current knowledge of breast cancer. *Biological research 50* (2017), 1–23.

[7] Alkabban, F. M., and Ferguson, T. Breast cancer - statpearls - ncbi bookshelf, Sep 2022.

[8] Andersson, I., Hildell, J., Muhlow, A., and Pettersson, H. Number of projections in mammography: influence on detection of breast disease. *American Journal of Roentgenology 130*, 2 (1978), 349–351.

[9] Arasu, V. A., Habel, L. A., Achacoso, N. S., Buist, D. S., Cord, J. B., Esserman, L. J., Hylton, N. M., Glymour, M. M., Kornak, J., Kushi, L. H., et al. Comparison of mammography ai algorithms with a clinical risk model for 5-year breast cancer risk prediction: An observational study. *Radiology 307*, 5 (2023), e222733.

[10] Asri, H., Mousannif, H., Al Moatassime, H., and Noel, T. Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Computer Science 83* (2016), 1064–1069.

[11] Bagdonavičius, V., Levulienė, R., and Nikulin, M. Goodness-of-fit criteria for the cox model from left truncated and right censored data. *Journal of Mathematical Sciences 167*, 4 (2010).

[12] Bassett, L. W., Hoyt, A. C., and Oshiro, T. Digital mammography: clinical image evaluation. *Radiologic Clinics 48*, 5 (2010), 903–915.

[13] Bednarek, A. K., Sahin, A., Brenner, A. J., Johnston, D. A., and Aldaz, C. M. Analysis of telomerase activity levels in breast cancer: positive detection at the in situ breast carcinoma stage. *Clinical cancer research: an official journal of the American Association for Cancer Research 3*, 1 (1997), 11–16.

[14] Bevers, T. B., Anderson, B. O., Bonaccio, E., Buys, S., Daly, M. B., Dempsey, P. J., Farrar, W. B., Fleming, I., Garber, J. E., Harris, R. E., et al. Breast cancer screening and diagnosis. *Journal of the National Comprehensive Cancer Network 7*, 10 (2009), 1060–1096.

[15] Bhagat, P., Choudhary, P., and Singh, K. M. Chapter 13 - a comparative study for brain tumor detection in mri images using texture features. In *Sensors for Health Monitoring*, N. Dey, J. Chaki, and R. Kumar, Eds., vol. 5 of *Advances in ubiquitous sensing applications for healthcare*. Academic Press, 2019, pp. 259–287.

[16] Binkley, J. M., Harris, S. R., Levangie, P. K., Pearl, M., Guglielmino, J., Kraus, V., and Rowden, D. Patient perspectives on breast cancer treatment side effects and the prospective surveillance model for physical rehabilitation for women with breast cancer. *Cancer 118*, S8 (2012), 2207–2216.

[17] Braitmaier, M., Kollhorst, B., Heinig, M., Langner, I., Czwikla, J., Heinze, F., Buschmann, L., Minnerup, H., García-Albéniz, X., Hense, H.-W., et al. Effectiveness of mammography screening on breast cancer mortality–a study protocol for emulation of target trials using german health claims data. *Clinical Epidemiology* (2022), 1293–1303.

[18] Breastcancer.org. Breast cancer facts and statistics. `https://www.breastcancer.org/facts-statistics`.

[19] Brentnall, A. R., and Cuzick, J. Risk models for breast cancer and their validation. *Statistical science: a review journal of the Institute of Mathematical Statistics 35*, 1 (2020), 14.

[20] Bürkner, P.-C., Gabry, J., and Vehtari, A. Efficient leave-one-out cross-validation for bayesian non-factorized normal and student-t models. *Computational Statistics 36*, 2 (2021), 1243–1261.

[21] Cai, J., Luo, J., Wang, S., and Yang, S. Feature selection in machine learning: A new perspective. *Neurocomputing 300* (2018), 70–79.

[22] Campello, V. M., Martín-Isla, C., Izquierdo, C., Guala, A., Palomares, J. F. R., Viladés, D., Descalzo, M. L., Karakas, M., Çavuş, E., Raisi-Estabragh, Z., et al. Minimising multi-centre radiomics variability through image normalisation: a pilot study. *Scientific Reports 12*, 1 (2022), 12532.

[23] Caseneuve, G., Valova, I., LeBlanc, N., and Thibodeau, M. Chest x-ray image preprocessing for disease classification. *Procedia Computer Science 192* (2021), 658–665.

[24] Celaya-Padilla, J. M., Rodriguez-Rojas, J., Galván-Tejada, J. I., Martínez-Torteya, A., Treviño, V., and Tamez-Peña, J. G. Bilateral image subtraction features for multivariate automated classification of breast cancer risk. In *Medical Imaging 2014: Computer-Aided Diagnosis* (2014), vol. 9035, SPIE, pp. 480–486.

[25] Chávarri-Guerra, Y., Villarreal-Garza, C., Liedke, P. E., Knaul, F., Mohar, A., Finkelstein, D. M., and Goss, P. E. Breast cancer in mexico: a growing challenge to health and the health system. *The Lancet Oncology 13*, 8 (2012), e335–e343.

[26] Choudhury, P. P., Wilcox, A. N., Brook, M. N., Zhang, Y., Ahearn, T., Orr, N., Coulson, P., Schoemaker, M. J., Jones, M. E., Gail, M. H., Swerdlow, A. J., Chatterjee, N., and Garcia-Closas, M. Comparative validation of breast cancer risk prediction models and projections for future risk stratification. *Journal of the National Cancer Institute 112* (2021), 278–285.

[27] Cintolo-Gonzalez, J. A., Braun, D., Blackford, A. L., Mazzola, E., Acar, A., Plichta, J. K., Griffin, M., and Hughes, K. S. Breast cancer risk models: a comprehensive overview of existing models, validation, and clinical applications. *Breast Cancer Research and Treatment 164* (7 2017), 263–284.

[28] Claus, E. B., Risch, N., and Thompson, W. D. Autosomal dominant inheritance of early-onset breast cancer. implications for risk prediction. *Cancer 73*, 3 (1994), 643–651.

[29] Cleator, S., Heller, W., and Coombes, R. C. Triple-negative breast cancer: therapeutic options. *The lancet oncology 8*, 3 (2007), 235–244.

[30] Corbex, M., Burton, R., and Sancho-Garnier, H. Breast cancer early detection methods for low and middle income countries, a review of the evidence. *The Breast 21*, 4 (2012), 428–434.

[31] Costantino, J. P., Gail, M. H., Pee, D., Anderson, S., Redmond, C. K., Benichou, J., and Wieand, H. S. Validation studies for models projecting the risk of invasive and total breast cancer incidence. *JNCI: Journal of the National Cancer Institute 91* (1999), 1541–1548.

[32] Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological) 34*, 2 (1972), 187–202.

[33] de Salud, S. Norma oficial mexicana nom-041-ssa2-2011, para la prevención, diagnóstico, tratamiento, control y vigilancia epidemiológica del cáncer de mama., Nov 2010. https://www.dof.gob.mx/nota_detalle.php?codigo=5194157&fecha=09%2F06%2F2011#gsc.tab=0.

[34] Eberl, M. M., Fox, C. H., Edge, S. B., Carter, C. A., and Mahoney, M. C. Bi-rads classification for management of abnormal mammograms. *The Journal of the American Board of Family Medicine 19*, 2 (2006), 161–164.

[35] England, P. H. Breast screening: Programme overview, Jun 2015. https://www.gov.uk/guidance/breast-screening-programme-overview.

[36] Eriksson, M., Czene, K., Pawitan, Y., Leifland, K., Darabi, H., and Hall, P. A clinical model for identifying the short-term risk of breast cancer. *Breast Cancer Research 19*, 1 (2017), 1–8.

[37] Eriksson, M., Czene, K., Vachon, C., Conant, E. F., and Hall, P. Long-term performance of an image-based short-term risk model for breast cancer. *Journal of Clinical Oncology* (2023), JCO–22.

[38] Eriksson, M., Destounis, S., Czene, K., Zeiberg, A., Day, R., Conant, E. F., Schilling, K., and Hall, P. A risk model for digital breast tomosynthesis to predict breast cancer and guide clinical care. *Science Translational Medicine 14*, 644 (2022), eabn3971.

[39] Euhus, D. M. Understanding mathematical models for breast cancer risk assessment and counseling. *The breast journal 7*, 4 (2001), 224–232.

[40] Faguet, G. B. A brief history of cancer: age-old milestones underlying our current knowledge database. *International journal of cancer 136*, 9 (2015), 2022–2036.

[41] Fan, J., Upadhye, S., and Worster, A. Understanding receiver operating characteristic (roc) curves. *Canadian Journal of Emergency Medicine 8*, 1 (2006), 19–20.

[42] Ferroni, P., Zanzotto, F. M., Riondino, S., Scarpato, N., Guadagni, F., and Roselli, M. Breast cancer prognosis using a machine learning approach. *Cancers 11*, 3 (2019), 328.

[43] Flach, P. Performance evaluation in machine learning: the good, the bad, the ugly, and the way forward. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2019), vol. 33, pp. 9808–9814.

[44] Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C., and Mulvihill, J. J. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *JNCI: Journal of the National Cancer Institute 81*, 24 (1989), 1879–1886.

[45] Ganesan, K., Acharya, U. R., Chua, C. K., Min, L. C., Abraham, K. T., and Ng, K.-H. Computer-aided breast cancer detection using mammograms: a review. *IEEE Reviews in biomedical engineering 6* (2012), 77–98.

[46] Gastounioti, A., Eriksson, M., Cohen, E. A., Mankowski, W., Pantalone, L., Ehsan, S., McCarthy, A. M., Kontos, D., Hall, P., and Conant, E. F. External validation of a mammography-derived ai-based risk model in a us breast cancer screening cohort of white and black women. *Cancers 14*, 19 (2022), 4803.

[47] Geroldinger, A., Lusa, L., Nold, M., and Heinze, G. Leave-one-out cross-validation, penalization, and differential bias of some prediction model performance measures—a simulation study. *Diagnostic and Prognostic Research 7*, 1 (2023), 1–11.

[48] Giampietro, R. R., Cabral, M. V. G., Lima, S. A. M., Weber, S. A. T., and dos Santos Nunes-Nogueira, V. Accuracy and effectiveness of mammography versus mammography and tomosynthesis for population-based breast cancer screening: a systematic review and meta-analysis. *Scientific reports 10*, 1 (2020), 1–10.

[49] Gilbert, F. J., and Pinker-Domenig, K. Diagnosis and staging of breast cancer: when and how to use mammography, tomosynthesis, ultrasound, contrast-enhanced mammography, and magnetic resonance imaging. *Diseases of the Chest, Breast, Heart and Vessels 2019-2022: Diagnostic and Interventional Imaging* (2019), 155–166.

[50] Gillies, R. J., and Schabath, M. B. Radiomics improves cancer screening and early detection. *Cancer Epidemiology Biomarkers and Prevention 29* (12 2020), 2556–2567.

[51] Gonzalez, R., and Woods, R. *Digital Image Processing*. Prentice Hall, 2008.

[52] Gøtzsche, P. C., and Jørgensen, K. J. Screening for breast cancer with mammography. *Cochrane database of systematic reviews*, 6 (2013).

[53] Grimm, L. J., Plichta, J. K., and Hwang, E. S. More than incremental: Harnessing machine learning to predict breast cancer risk, 2022.

[54] Group, E. B. C. T. C., et al. Effects of radiotherapy and of differences in the extent of surgery for early breast cancer on local recurrence and 15-year survival: an overview of the randomised trials. *The Lancet 366*, 9503 (2005), 2087–2106.

[55] Gu, G., Dustin, D., and Fuqua, S. A. Targeted therapy for breast cancer and molecular mechanisms of resistance to treatment. *Current opinion in pharmacology 31* (2016), 97–103.

[56] Handelman, G. S., Kok, H. K., Chandra, R. V., Razavi, A. H., Huang, S., Brooks, M., Lee, M. J., and Asadi, H. Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods. *American Journal of Roentgenology 212*, 1 (2019), 38–43.

[57] Haralick, R. M., Shanmugam, K., and Dinstein, I. H. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, 6 (1973), 610–621.

[58] Hassan, M., Suhail Shaikh, M., and Jatoi, M. A. Image quality measurement-based comparative analysis of illumination compensation methods for face image normalization. *Multimedia Systems* (2022), 1–10.

[59] Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer, 2009.

[60] Heddson, B., Rönnow, K., Olsson, M., and Miller, D. Digital versus screen-film mammography: a retrospective comparison in a population-based screening program. *European journal of radiology 64*, 3 (2007), 419–425.

[61] Himes, D. O., Root, A. E., Gammon, A., and Luthy, K. E. Breast cancer risk assessment: calculating lifetime risk using the tyrer-cuzick model. *The Journal for Nurse Practitioners 12*, 9 (2016), 581–592.

[62] Hossin, M., and Sulaiman, M. N. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process 5*, 2 (2015), 1.

[63] Houssami, N. Overdiagnosis of breast cancer in population screening: does it make breast screening worthless? *Cancer biology & medicine 14*, 1 (2017), 1.

[64] Independ, U. Panel on breast cancer screening. the benefits and harms of breast cancer screening: an independent review. *Lancet 380*, 9855 (2012), 1778–86.

[65] Iranmakani, S., Mortezazadeh, T., Sajadian, F., Ghaziani, M. F., Ghafari, A., Khezerloo, D., and Musa, A. E. A review of various modalities in breast imaging: technical aspects and clinical outcomes. *Egyptian Journal of Radiology and Nuclear Medicine 51*, 1 (2020), 1–22.

[66] Jacquillat, C., Weil, M., Baillet, F., Borel, C., Auclerc, G., De Maublanc, M., Housset, M., Forget, G., Thill, L., Soubrane, C., et al. Results of neoadjuvant chemotherapy and radiation therapy in the breast-conserving treatment of 250 patients with all stages of infiltrative breast cancer. *Cancer 66*, 1 (1990), 119–129.

[67] Jung, Y. M. *Data Analysis in Quantitative Research*. Springer Singapore, Singapore, 2019, pp. 955–969.

[68] Ke-Chen, S., Yun-Hui, Y., Wen-Hui, C., and Zhang, X. Research and perspective on local binary pattern. *Acta Automatica Sinica 39*, 6 (2013), 730–744.

[69] Kellen, E., Vansant, G., Christiaens, M.-R., Neven, P., and Van Limbergen, E. Lifestyle changes and breast cancer prognosis: a review. *Breast cancer research and treatment 114* (2009), 13–22.

[70] Khan, A. I., and Al-Habsi, S. Machine learning in computer vision. *Procedia Computer Science 167* (2020), 1444–1451.

[71] Kim, G., and Bahl, M. Assessing risk of breast cancer: a review of risk prediction models. *Journal of breast imaging 3*, 2 (2021), 144–155.

[72] King, R. D., Orhobor, O. I., and Taylor, C. C. Cross-validation is safe to use. *Nature Machine Intelligence 3*, 4 (2021), 276–276.

[73] Kleinbaum, D. G., and Klein, M. *Survival analysis a self-learning text*. Springer, 1996.

[74] Kocher, M., Ruge, M. I., Galldiks, N., and Lohmann, P. Applications of radiomics and machine learning for radiotherapy of malignant brain tumors. *Strahlentherapie und Onkologie 196* (2020), 856–867.

[75] Kohavi, R., et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (1995), vol. 14, Montreal, Canada, pp. 1137–1145.

[76] Kontos, D., Winham, S. J., Oustimov, A., Pantalone, L., Hsieh, M.-K., Gastounioti, A., Whaley, D. H., Hruska, C. B., Kerlikowske, K., Brandt, K., et al. Radiomic phenotypes of mammographic parenchymal complexity: toward augmenting breast density in breast cancer risk assessment. *Radiology 290*, 1 (2019), 41–49.

[77] Kori, S. An overview: Several causes of breast cancer. *Epidemol. Int. J 2* (2018), 000107.

[78] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal 13* (2015), 8–17.

[79] Kumar, V., Gu, Y., Basu, S., Berglund, A., Eschrich, S. A., Schabath, M. B., Forster, K., Aerts, H. J., Dekker, A., Fenstermacher, D., et al. Radiomics: the process and the challenges. *Magnetic resonance imaging 30*, 9 (2012), 1234–1248.

[80] Kurian, A. W., Hughes, E., Simmons, T., Bernhisel, R., Probst, B., Meek, S., Caswell-Jin, J. L., John, E. M., Lanchbury, J. S., Slavin, T. P., et al. Performance of the ibis/tyrer-cuzick model of breast cancer risk by race and ethnicity in the women's health initiative. *Cancer 127*, 20 (2021), 3742–3750.

[81] Lauby-Secretan, B., Scoccianti, C., Loomis, D., Benbrahim-Tallaa, L., Bouvard, V., Bianchini, F., and Straif, K. Breast-cancer screening—viewpoint of the iarc working group. *New England journal of medicine 372*, 24 (2015), 2353–2358.

[82] Lee, H., Kim, J., Park, E., Kim, M., Kim, T., and Kooi, T. Enhancing breast cancer risk prediction by incorporating prior images. *arXiv preprint arXiv:2303.15699* (2023).

[83] Liu, Y., He, M., Zuo, W.-J., Hao, S., Wang, Z.-H., and Shao, Z.-M. Tumor size still impacts prognosis in breast cancer with extensive nodal involvement. *Frontiers in Oncology 11* (2021), 585613.

[84] Liu, Y.-Q., Wang, C., and Zhang, L. Decision tree based predictive models for breast cancer survivability on imbalanced data. In *2009 3rd international conference on bioinformatics and biomedical engineering* (2009), IEEE, pp. 1–4.

[85] Løberg, M., Lousdal, M. L., Bretthauer, M., and Kalager, M. Benefits and harms of mammography screening. *Breast Cancer Research 17*, 1 (2015), 1–12.

[86] Louro, J., Posso, M., Boon, M. H., Román, M., Domingo, L., Castells, X., and Sala, M. A systematic review and quality assessment of individualised breast cancer risk prediction models. *British journal of cancer 121*, 1 (2019), 76–85.

[87] Lukasiewicz, S., Czeczelewski, M., Forma, A., Baj, J., Sitarz, R., and Stanis lawek, A. Breast cancer—epidemiology, risk factors, classification, prognostic markers, and current treatment strategies—an updated review. *Cancers 13*, 17 (2021), 4287.

[88] Lukong, K. E. Understanding breast cancer–the long and winding road. *BBA clinical 7* (2017), 64–77.

[89] Mahesh, B. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet] 9* (2020), 381–386.

[90] Mali, S. A., Ibrahim, A., Woodruff, H. C., Andrearczyk, V., Müller, H., Primakov, S., Salahuddin, Z., Chatterjee, A., and Lambin, P. Making radiomics more reproducible across scanner and imaging protocol variations: a review of harmonization methods. *Journal of personalized medicine 11*, 9 (2021), 842.

[91] Manganaro, L., Nicolino, G. M., Dolciami, M., Martorana, F., Stathis, A., Colombo, I., and Rizzo, S. Radiomics in cervical and endometrial cancer. *The British Journal of Radiology 94*, 1125 (2021), 20201314.

[92] Masoudi, S., Harmon, S. A., Mehralivand, S., Walker, S. M., Raviprakash, H., Bagci, U., Choyke, P. L., and Turkbey, B. Quick guide on radiology image pre-processing for deep learning applications in prostate cancer research. *Journal of Medical Imaging 8*, 1 (2021), 010901–010901.

[93] Mendoza, J. Number of breast cancer deaths among women in mexico from 2010 to 2021, Mar 2023. `https://www.statista.com/statistics/999090/mexico-breast-cancer-mortality/`.

[94] Metz, C. E. Basic principles of roc analysis. In *Seminars in nuclear medicine* (1978), vol. 8, Elsevier, pp. 283–298.

[95] Miller, K. D., Goding Sauer, A., Ortiz, A. P., Fedewa, S. A., Pinheiro, P. S., Tortolero-Luna, G., Martinez-Tyson, D., Jemal, A., and Siegel, R. L. Cancer statistics for hispanics/latinos, 2018. *CA: a cancer journal for clinicians 68*, 6 (2018), 425–445.

[96] Ming, C., Viassolo, V., Probst-Hensch, N., Dinov, I. D., Chappuis, P. O., and Katapodi, M. C. Machine learning-based lifetime breast cancer risk reclassification compared with the boadicea model: impact on screening recommendations. *British journal of cancer 123*, 5 (2020), 860–867.

[97] Monticciolo, D. L., Newell, M. S., Moy, L., Niell, B., Monsees, B., and Sickles, E. A. Breast cancer screening in women at higher-than-average risk: Recommendations from the acr. *Journal of the American College of Radiology 15* (3 2018), 408–414.

[98] Moran, M. S., Schnitt, S. J., Giuliano, A. E., Harris, J. R., Khan, S. A., Horton, J., Klimberg, S., Chavez-MacGregor, M., Freedman, G., Houssami, N., et al. Society of surgical oncology–american society for radiation oncology consensus guideline on margins for breast-conserving surgery with whole-breast irradiation in stages i and ii invasive breast cancer. *International Journal of Radiation Oncology* Biology* Physics 88*, 3 (2014), 553–564.

[99] Moulder, S., and Hortobagyi, G. Advances in the treatment of breast cancer. *Clinical Pharmacology & Therapeutics 83*, 1 (2008), 26–36.

[100] Moulin, P. *Multiscale Image Decompositions and Wavelets*. Elsevier, 1 2009.

[101] Naji, M. A., El Filali, S., Aarika, K., Benlahmar, E. H., Abdelouhahid, R. A., and Debauche, O. Machine learning algorithms for breast cancer prediction and diagnosis. *Procedia Computer Science 191* (2021), 487–492.

[102] Nees, A. V. Digital mammography: Are there advantages in screening for breast cancer? *Academic Radiology 15*, 4 (2008), 401–407.

[103] Nelson, H. D., Fu, R., Cantor, A., Pappas, M., Daeges, M., and Humphrey, L. Effectiveness of breast cancer screening: systematic review and meta-analysis to update the 2009 us preventive services task force recommendation. *Annals of internal medicine 164*, 4 (2016), 244–255.

[104] Neuman, H. B., Morrogh, M., Gonen, M., Van Zee, K. J., Morrow, M., and King, T. A. Stage iv breast cancer in the era of targeted therapy: does surgery of the primary tumor matter? *Cancer: Interdisciplinary International Journal of the American Cancer Society 116*, 5 (2010), 1226–1233.

[105] on the Evaluation of Cancer-Preventive Strategies, I. W. G., and Organization, W. H. *Breast Cancer Screening.*, vol. 15. IARC HANDBOOKS OF CANCER PREVENTION, 2016.

[106] Organization, W. H. Breast cancer, Mar 2021. `https://www.who.int/news-room/fact-sheets/detail/breast-cancer`.

[107] Orozco Sanchez, J. A. *Machine Learning and Cox Based Benchmarking Tool: Exploration of Survival Models Associated with Chronic Degenerative Diseases*. PhD thesis, Tecnologico de Monterrey, 2020.

[108] Oyelade, O. N., and Ezugwu, A. E. A novel wavelet decomposition and transformation convolutional neural network with data augmentation for breast cancer detection using digital mammogram. *Scientific Reports 12*, 1 (2022), 5913.

[109] Palesh, O., Scheiber, C., Kesler, S., Mustian, K., Koopman, C., and Schapira, L. Management of side effects during and post-treatment in breast cancer survivors. *The breast journal 24*, 2 (2018), 167–175.

[110] Parida, P., and Bhoi, N. Wavelet based transition region extraction for image segmentation. *Future Computing and Informatics Journal 2*, 2 (2017), 65–78.

[111] Pham, D. L., Xu, C., and Prince, J. L. Current methods in medical image segmentation. *Annual review of biomedical engineering 2*, 1 (2000), 315–337.

[112] Pietikäinen, M., Hadid, A., Zhao, G., and Ahonen, T. *Computer vision using local binary patterns*, vol. 40. Springer Science & Business Media, 2011.

[113] Ping Tian, D., et al. A review on image feature extraction and representation techniques. *International Journal of Multimedia and Ubiquitous Engineering 8*, 4 (2013), 385–396.

[114] Polo, T. C. F., and Miot, H. A. Use of roc curves in clinical and experimental studies, 2020.

[115] Refaeilzadeh, P., Tang, L., and Liu, H. *Cross-Validation*. Springer US, Boston, MA, 2009, pp. 532–538.

[116] Reynoso-Noverón, N., Villarreal-Garza, C., Soto-Perez-de Celis, E., Arce-Salinas, C., Matus-Santos, J., Ramírez-Ugalde, M. T., Alvarado-Miranda, A., Cabrera-Galeana, P., Meneses-García, A., Lara-Medina, F., et al. Clinical and epidemiological profile of breast cancer in mexico: Results of the seguro popular. *Journal of global oncology 3*, 6 (2017), 757–764.

[117] Richie, R. C., and Swanson, J. O. Breast cancer: a review of the literature. *JOURNAL OF INSURANCE MEDICINE-NEW YORK THEN DENVER– 35*, 2 (2003), 85–101.

[118] Sahoo, P. K., Soltani, S., and Wong, A. K. A survey of thresholding techniques. *Computer vision, graphics, and image processing 41*, 2 (1988), 233–260.

[119] Sahoo, S. S., Kobow, K., Zhang, J., Buchhalter, J., Dayyani, M., Upadhyaya, D. P., Prantzalos, K., Bhattacharjee, M., Blumcke, I., Wiebe, S., et al. Ontology-based feature engineering in machine learning workflows for heterogeneous epilepsy patient records. *Scientific Reports 12*, 1 (2022), 19430.

[120] Sandilands, D. D. *Univariate Analysis*. Springer Netherlands, Dordrecht, 2014, pp. 6815–6817.

[121] Sebastian V, B., Unnikrishnan, A., and Balakrishnan, K. Gray level co-occurrence matrices: generalisation and some new features. *arXiv preprint arXiv:1205.4831* (2012).

[122] Segal, R., Evans, W., Johnson, D., Smith, J., Colletta, S., Gayton, J., Woodard, S., Wells, G., and Reid, R. Structured exercise improves physical functioning in women with stages i and ii breast cancer: results of a randomized controlled trial. *Journal of clinical oncology 19*, 3 (2001), 657–665.

[123] Shah, R., Rosso, K., and Nathanson, S. D. Pathogenesis, prevention, diagnosis and treatment of breast cancer. *World journal of clinical oncology 5*, 3 (2014), 283.

[124] Sickles, E., Weber, W., Galvin, H., Ominsky, S., and Sollitto, R. Baseline screening mammography: one vs two views per breast. *American Journal of Roentgenology 147*, 6 (1986), 1149–1153.

[125] Skaane, P. Studies comparing screen-film mammography and full-field digital mammography in breast cancer screening: updated review. *Acta radiologica 50*, 1 (2009), 3–14.

[126] Smith-Bindman, R., Miglioretti, D. L., Lurie, N., Abraham, L., Barbash, R. B., Strzelczyk, J., Dignan, M., Barlow, W. E., Beasley, C. M., and Kerlikowske, K. Does utilization of screening mammography explain racial and ethnic differences in breast cancer? *Annals of internal medicine 144*, 8 (2006), 541–553.

[127] Society, A. C. American cancer society guidelines for the early detection of cancer. `https://www.cancer.org/cancer/screening/american-cancer-society-guidelines-for-the-early-detection-of-cancer.html`.

[128] Sonka, M., Hlavac, V., Boyle, R., Sonka, M., Hlavac, V., and Boyle, R. Image pre-processing. *Image processing, analysis and machine vision* (1993), 56–111.

[129] Spak, D. A., Plaxco, J., Santiago, L., Dryden, M., and Dogan, B. Bi-rads® fifth edition: A summary of changes. *Diagnostic and interventional imaging 98*, 3 (2017), 179–190.

[130] Stang, A., Kääb-Sanyal, V., Hense, H.-W., Becker, N., and Kuss, O. Effect of mammography screening on surgical treatment for breast cancer: a nationwide analysis of hospitalization rates in germany 2005–2009. *European journal of epidemiology 28* (2013), 689–696.

[131] Suhrke, P., Mæhlen, J., Schlichting, E., Jørgensen, K. J., Gøtzsche, P. C., and Zahl, P.-H. Effect of mammography screening on surgical treatment for breast cancer in norway: comparative analysis of cancer registry data. *Bmj 343* (2011).

[132] Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians 71*, 3 (2021), 209–249.

[133] Sweeney, R.-J. I., Lewis, S. J., Hogg, P., and McEntee, M. F. A review of mammographic positioning image quality criteria for the craniocaudal projection. *The British journal of radiology 91*, 1082 (2018), 20170611.

[134] Tabachnick, B. G., Fidell, L. S., and Ullman, J. B. *Using multivariate statistics*, vol. 6. pearson Boston, MA, 2013.

[135] Tamez-Pena, J. Benchmarking binary classifiers algorithms.

[136] Tamez-P'ena, J. A., and De la Garza, J. M. Feature selection and the bswims method. In *Proceedings of the 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (2018), IEEE, pp. 2899–2904.

[137] Tamez-Peña, J.-G., Rodriguez-Rojas, J.-A., Gomez-Rueda, H., Celaya-Padilla, J.-M., Rivera-Prieto, R.-A., Palacios-Corona, R., Garza-Montemayor, M., Cardona-Huerta, S., and Treviño, V. Radiogenomics analysis identifies correlations of digital mammography with clinical molecular signatures in breast cancer. *PloS one 13*, 3 (2018), e0193871.

[138] Tibshirani, R. Regression Shrinkage and Selection via The Lasso: A Retrospective. *Journal of the Royal Statistical Society Series B: Statistical Methodology 73*, 3 (2011), 273–282.

[139] Tice, J. A., Cummings, S. R., Smith-Bindman, R., Ichikawa, L., Barlow, W. E., and Kerlikowske, K. Using clinical factors and mammographic breast density to estimate breast cancer risk: development and validation of a new predictive model. *Annals of internal medicine 148*, 5 (2008), 337–347.

[140] Tyrer, J., Duffy, S. W., and Cuzick, J. A breast cancer prediction model incorporating familial and personal risk factors. *Statistics in medicine 23*, 7 (2004), 1111–1130.

[141] van Timmeren, J. E., Cester, D., Tanadini-Lang, S., Alkadhi, H., and Baessler, B. Radiomics in medical imaging—"how-to" guide and critical reflection. *Insights into Imaging 11* (12 2020).

[142] Venkataraman, V., Browning, T., Pedrosa, I., Abbara, S., Fetzer, D., Toomay, S., and Peshock, R. M. Implementing shared, standardized imaging protocols to improve cross-enterprise workflow and quality. *Journal of Digital Imaging 32* (2019), 880–887.

[143] Verdonck, T., Baesens, B., Óskarsdóttir, M., and vanden Broucke, S. Special issue on feature engineering editorial. *Machine Learning* (2021), 1–12.

[144] Vikhe, P., and Thool, V. Mass detection in mammographic images using wavelet processing and adaptive threshold technique. *Journal of medical systems 40* (2016), 1–16.

[145] Wang, H., Ma, C., and Zhou, L. A brief review of machine learning and its application. In *2009 international conference on information engineering and computer science* (2009), IEEE, pp. 1–4.

[146] Wang, X., Huang, Y., Li, L., Dai, H., Song, F., and Chen, K. Assessment of performance of the gail model for predicting breast cancer risk: A systematic review and meta-analysis with trial sequential analysis. *Breast Cancer Research 20* (3 2018).

[147] Wang, X., and Paliwal, K. K. Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition. *Pattern recognition 36*, 10 (2003), 2429–2439.

[148] Weiderpass, E., and Stewart, B. World cancer report: Cancer research for cancer prevention, 2020.

[149] Winters, S., Alomari, A., Shokar, G., Martin, C., Dwivedi, A., and Shokar, N. K. Breast cancer screening outcomes among mexican-origin hispanic women participating in a breast cancer screening program. *Preventive Medicine Reports 24* (2021), 101561.

[150] Wolfe, J. N. A study of breast parenchyma by mammography in the normal woman and those with benign and malignant disease. *Radiology 89*, 2 (1967), 201–205.

[151] Yala, A., Lehman, C., Schuster, T., Portnoi, T., and Barzilay, R. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology 292* (2019), 60–66.

[152] Yala, A., Mikhael, P. G., Strand, F., Lin, G., Smith, K., Wan, Y.-L., Lamb, L., Hughes, K., Lehman, C., and Barzilay, R. Toward robust mammography-based models for breast cancer risk. *Science Translational Medicine 13*, 578 (2021), eaba4373.

[153] Yedjou, C. G., Sims, J. N., Miele, L., Noubissi, F., Lowe, L., Fonseca, D. D., Alo, R. A., Payton, M., and Tchounwou, P. B. Health and racial disparity in breast cancer. *Breast cancer metastasis and drug resistance: Challenges and progress* (2019), 31–49.

[154] Zanaty, E., and Ghoniemy, S. Medical image segmentation techniques: an overview. *International Journal of informatics and medical data processing 1*, 1 (2016), 16–37.

[155] Zhang, D., et al. *Fundamentals of image data mining*. Springer, 2019.

[156] Zhang, X., Rice, M., Tworoger, S. S., Rosner, B. A., Eliassen, A. H., Tamimi, R. M., Joshi, A. D., Lindstrom, S., Qian, J., Colditz, G. A., et al. Addition of a polygenic risk score, mammographic density, and endogenous hormones to existing breast cancer risk prediction models: A nested case–control study. *PLoS medicine 15*, 9 (2018), e1002644.

[157] Zheng, A., and Casari, A. *Feature engineering for machine learning: principles and techniques for data scientists*. " O'Reilly Media, Inc.", 2018.

[158] Zhong, X., Gallagher, B., Eves, K., Robertson, E., Mundhenk, T. N., and Han, T. Y.-J. A study of real-world micrograph data quality and machine learning model robustness. *npj Computational Materials 7*, 1 (2021), 161.