POLITECNICO DI TORINO

Master of Science in Biomedical Engineering
Department of Electronics and Telecommunication

Master of Science Thesis

Computer-Aided Diagnosis of Breast Masses using Shape and Margin Radiomic Descriptors and Neural Networks in Dedicated Breast CT Imaging

Academic Supervisors
Prof. Filippo MOLINARI
Prof. Luca MAINARDI

Candidate
Domenico Ruben PANGALLO

Company Supervisor
Ing. Marco CABALLO
Radboud Imaging Research Center
Radboudumc, Nijmegen, The Netherlands

ACADEMIC YEAR 2018-2019

For my beloved parents,
you encouraged me to pursue my dreams.

Acknowledgements

First of all, I would like to thank all those people who contributed to the realization of this thesis, which made me grow not only from the academic perspective but also from the human and personal points of view. Therefore, I am grateful to Professor Filippo Molinari for believing in the project from the start and giving me confidence over my six months abroad. Sincere appreciation to Professor Luca Mainardi from Politecnico di Milano for the patience shown in his thesis co-supervision. And thanks to you, Marco, for allowing me to work with you and having always been close to me for all these months. You gave me your time, support, and knowledge, considering me as the others in the Lab: you were not just a supervisor, but also a true friend.

Secondly but not less important, I would like to thank my parents, who allowed me to make this incredible experience. These few words cannot express all the gratitude and appreciation I feel about you. This achievement is for you, who support me incessantly and always push me to do my best.

A special thanks also to all my family, especially, to you, Gianluca, for being like a brother to me. We shared every moment of our lives, and we helped each other out in those months when we were both involved in our thesis. You are always there for me in every circumstance, despite the distance between us. Che bellezza!

And last but not least, I dedicate this thesis to all my friends, those of a lifetime, those known in these years of studies, those with whom I lived the ASP experience, those with whom we shared Knowai, those I met in my Dutch adventure: thank you for sharing moments of happiness with me and always leaving me something to learn from. They were five years of fun, sacrifices, enthusiasm, difficulties, but, in one way or another, I have always felt your support and your presence.

Domenico Ruben Pangallo

Prima di tutto, vorrei esprimere il mio ringraziamento a tutti coloro che hanno contribuito alla realizzazione di questa tesi, che mi ha permesso di crescere non solo da un punto di vista accademico, ma anche umano e personale. Dunque, grazie al Professor Filippo Molinari per aver creduto da subito nel progetto e avermi dato fiducia nell'arco dei miei mesi all'estero. Sincero apprezzamento al Professor Luca Mainardi del Politecnico di Milano per la pazienza mostrata nella sua co-supervisione della tesi. E grazie a te, Marco, per avermi dato l'opportunità di lavorare con te ed essermi sempre stato vicino durante tutti questi mesi. Tu mi hai dato il tuo tempo, il tuo supporto, la tua conoscenza, mettendomi alla pari di tutti gli altri del laboratorio: non sei stato solamente un tutor, ma un vero amico.

In secondo luogo ma non meno importanti, vorrei ringraziare i miei genitori, che mi hanno permesso di compiere questa straordinaria esperienza. Queste poche parole non possono esprimere tutta la gratitudine e l'apprezzamento che provo nei vostri confronti. Questo traguardo è per voi, che mi sostenete incessantemente e mi spingete sempre a dare il massimo.

E un ringraziamento speciale va anche a tutta la mia famiglia e in particolare a te, Gianluca, che sei come un fratello per me. Abbiamo condiviso ogni attimo delle nostre vite e ci siamo dati manforte in questi mesi in cui entrambi eravamo coinvolti nello svolgimento della tesi. Tu ci sei sempre per me in ogni circostanza, nonostante la distanza che ci separa. Che bellezza!

Infine, dedico questa tesi a tutti i miei amici, a quelli di una vita, a quelli conosciuti in questi anni di studi al Politecnico, a quelli con cui ho vissuto l'esperienza ASP, a quelli con cui abbiamo condiviso Knowai, a quelli incontrati nella mia avventura olandese: grazie per aver condiviso con me attimi di felicità e avermi sempre lasciato qualcosa da cui imparare. Sono stati 5 anni di divertimenti, sacrifici, entusiasmo, momenti di difficoltà, ma in un modo o nell'altro ho sempre percepito il vostro sostegno e la vostra presenza.

Domenico Ruben Pangallo

Abstract

Objective:

The goal of this Master Thesis is to design a Computer-Aided Diagnosis (CADx) system for breast mass-like lesion classification in Dedicated Breast CT (DBCT) images, using a quantitative radiomics approach based on newly developed shape and margin imaging biomarkers and a multi-layer perceptron artificial neural network (ANN). The clinical motivation behind it is to reduce the number of negative and unnecessary breast biopsies, which constitutes more than 70% of the overall biopsies performed. This project was carried out at the Advanced X-ray Tomographic Imaging (AXTI) Laboratory, Department of Radiology and Nuclear Medicine, Radboud University Medical Center (Nijmegen, The Netherlands).

Methods:

A traditional radiomic pipeline was implemented. Therefore, starting from DBCT images and their manual segmentation, the main phases performed were the *image cropping* to obtain the patches containing the breast lesions, the *data augmentation* process to increase the number of available patches, the *feature extraction* based on shape and margin descriptors, and the implementation of the Machine Learning (ML) diagnostic model for the classification of benign and malignant breast masses. The images used in this study were acquired with the new DBCT imaging modality and consisted of 74 breast lesions (54 benign and 20 malignant).

As concerns the breast masses analysis, this work was focused on the development, validation, and implementation of new descriptors that quantify the tumor properties in terms of morphology (*shape and contour*) and margin appearance (*border heterogeneity and infiltration degree*). The final feature set contained 158 features, some of which were already included in previously published studies, others were newly designed and proposed in this thesis.

As regards the classification task, this work involved a study related to the ANN architecture in terms of the number of hidden layers and neurons (*ANN tuning*), the number of features to consider (*feature selection*), and the number of samples to be used during the network training (*dataset balancing*) in order to correctly predict the breast lesion diagnosis.

Results:

The final ANN model resulted in a sensitivity of 0.79, a specificity of 0.90, an F1-Score of 0.79, an accuracy of 0.86, and an AUC_{ROC} of 0.91. It contained 56 input features (feature selection using *Random Forest*), two hidden layers of 10 neurons each with a hyperbolic tangent sigmoid activation function, and one output layer with a logistic sigmoid one. The dataset used for its training was balanced by undersampling the class of benign lesions (majority class). In particular, *dataset balancing* made it possible to significantly increase the classification performance, compared to ANN models based on the initial imbalanced dataset (benign to malignant ratio of 3:1). Instead, *feature selection* showed that its contribution was minimal both in the presence of imbalanced and balanced datasets, so it did not constitute a critical step, underlying the importance of these features in the discrimination between benign and malignant masses.

Conclusion:

The inclusion of radiomic features that assess morphology, margin and peritumoral compartments of breast lesions in the development of a CADx system for DBCT could provide high classification performance. Therefore, their combination with the well-known and traditional texture features could lead to a radiomic analysis aimed at quantifying all most distinctive characteristics of breast masses, resulting in a significant diagnostic decision support.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Breast Cancer

1.1.1 Epidemiology

Breast cancer was the second most common cancer type worldwide, affecting more than 2 million people (11.6% of the total number of new cancer cases), and the fifth most common cause of cancer death with a total of 627000 cases (6.6%) in 2018.
By analyzing only women, breast cancer was the most common type of cancer (25.4%) and the leading cause of cancer death worldwide (15%) in 2018 [1].

Breast cancer survival rate ranges widely because the worldwide situation shows how it varies from over 85% in developed and high-income countries to around 40% in underdeveloped and low-income ones [2].
These statistics demonstrate the importance of early detection programs started in the major world countries to prevent women from having late-stage cancers, more difficult to be treated.
Considering the United States scenario [3], it has been shown that the 5-year Relative Survival Rate (RSR), which describes the percentage of people who are alive at least 5 years after the diagnosis, is around 90% for women with invasive breast cancer (62% of diagnosed cases). This parameter varies greatly with the stage of cancer: it drops to 85% and 27% if cancer has spread to a nearby region (e.g. lymph nodes, stage III) or metastasized to a distant part of the body (e.g. liver, stage IV), respectively [4]. Therefore, these statistics further underline the importance of efficient screening and diagnosis programs to determine in a reliable and fast way which type of cancer a person has.

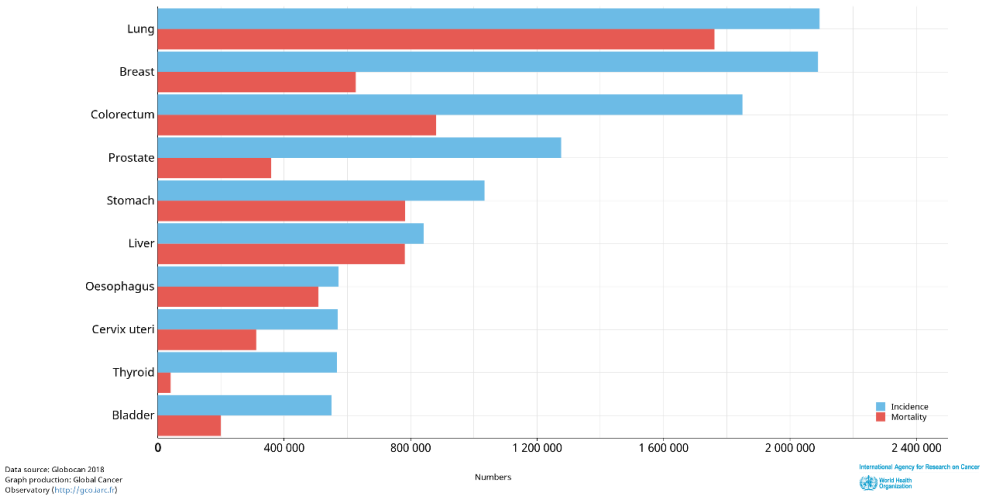


Figure 1. Incidence and Mortality of cancer in the world in 2018. It shows that breast cancer is the second most occurring cancer worldwide, but it is the fifth cancer in terms of deaths (ranking behind lung, colorectum, stomach and liver). Retrieved from [1].

1.1.2 Breast Tumor Types

Breast cancer indicates a group of diseases that occur when abnormal cells of the breast grow out of control and excess as compared to surrounding healthy tissues: this lump of cancer cells can subsequently increase in number (unchecked reproduction), spreading to the other healthy areas of the breast (and, at worst, of the body). The build-up of these mammary cells raised irregularly due to the alteration of their cellular vitality, involves the birth of a mass of tissue which is called **neoplasm** or **tumor**.

Figure 2 shows the anatomy of the breast, that is composed of different types of tissue:

- *Glandular* tissue, which comprehends the lobules and ducts;
- *Fibrous* tissue, which is the main element of ligaments (and of scar tissues);
- *Adipose* (or *Fatty*) tissue, which is the least dense and fills in the spaces between the two previously mentioned tissues.

One of the distinguishing characteristics of the breast is the density, which is described as the quantity of fibrous and glandular tissue present within the breast volume. The Breast Imaging-Reporting and Data System (BI-RADS), a guide that provides standardized terminology and assessment categories in breast imaging, contains a section dedicated to "Breast Composition Categories": this classification in four categories (Fatty, Scattered fibroglandular density, Heterogeneously dense, Extremely dense) is really important because breast density affects the quality of the image, and the detection and diagnosis processes. Indeed, as will be discussed below, breast density has a masking effect on tumors, which in most cases are located in the fibroglandular component (i.e. lobules and duct walls) [5]. Moreover, breast density is related to the risk of developing breast cancer because the more glandular tissue the patient has, the more likely cancer will develop [6].

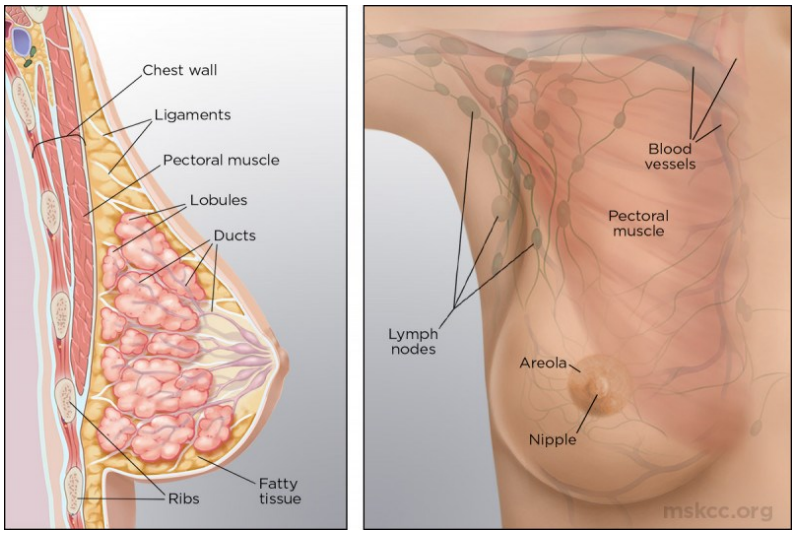


Figure 2. Anatomy of the breast. Retrieved from [7].

The main point to be stressed is that breast tumors can be of two types, “benign” or “malignant”.

The benign tumors are non-invasive and non-cancerous, but they should be monitored because if they grew or modified their shape, it would be necessary to surgically remove them before they cause pain or complications.
Types of benign breast masses include cysts, fibroadenomas, and central intraductal papillomas (or atypical papillomas).
The cysts are fluid-filled sacs which can cause pain, but do not increase the cancer risks: about 25% of all breast masses belong to this class [8].
The most common benign breast tumor is the fibroadenoma, which appears as a round and well-defined shape lump and does not usually cause tenderness: the name put together the fibrous and glandular nature of its tissue composition.
Lastly, atypical papillomas are small growths that develop in the breast ducts and consist of a mix of glandular and fibrovascular tissues. Their name refers to their central position, as they grow in the proximity of the nipple.

Malignant tumors, or cancers, can damage the surrounding healthy tissues. When dealing with breast cancers, one of the common practices is to analyze the aggressiveness stage of the tumor, defining its grade on a low (well-differentiated), intermediate (moderately differentiate), high (poorly differentiated) scale [9].
They can be divided into two large families: non-invasive and invasive.
Among the first group, there is the Ductal Carcinoma in Situ (DCIS), which is found in the duct cells and is not so spread through the other parts of the breast. It is considered an early-stage cancer lesion, and, in some cases, it eventually proceeds to the next invasive cancer stage [10].
Among the second group, there are the Invasive Ductal Carcinoma (IDC) and the Invasive Lobular Carcinoma (ILC). The former is the infiltrating ductal carcinoma and, differently from DCIS, spreads to the rest of the breast and of the body: it represents around 80% of all breast cancers. The latter is less common than previous cancer (around 10%) and develops from the lobule cells where the milk is produced [8]. Aside from the types previously described, breast cancer can manifest in additional neoplastic forms, whose prevalence is significantly lower and, therefore, whose description goes beyond the scope of this thesis.

1.1.3 Main Differences between Benign and Malignant Masses

Benign and malignant breast mass-like tumors exhibit a set of morphological connotations from whose analysis and characterization it is possible to try to distinguish them and assess their degree of malignancy. In particular, four different morphological aspects can be investigated: shape, boundary irregularities, degree of spiculation and degree of infiltration in the tumor periphery.
Indeed, the BI-RADS guide uses these descriptors as qualitative criteria for breast masses examination, subdividing them into categories to standardize the vocabulary and to improve the communication for all those involved in the detection, diagnosis, and treatment of these lesions (e.g. radiologists, pathologists).

MAMMOGRAPHY		
Breast composition	a. The breasts are almost entirely fatty b. There are scattered areas of fibroglandular density c. The breasts are heterogeneously dense, which may obscure small masses d. The breasts are extremely dense, which lowers the sensitivity of mammography	
Masses	Shape	Oval Round Irregular
	Margin	Circumscribed Obscured Microlobulated Indistinct Spiculated
	Density	High density Equal density Low density Fat-containing

ULTRASOUND		
Tissue composition *(screening only)*	a. Homogeneous background echotexture – fat b. Homogeneous background echotexture – fibroglandular c. Heterogeneous background echotexture	
Masses	Shape	Oval Round Irregular
	Orientation	Parallel Not parallel
	Margin	Circumscribed Not circumscribed - Indistinct - Angular - Microlobulated - Spiculated
	Echo pattern	Anechoic

Figure 3. Mammography and Ultrasound chapters of the ACR BI-RADS Fifth Edition guide that show how shape and margin are two of the descriptors used for describing and classifying solid breast masses. Retrieved from [11].

Regarding benign masses, they are usually characterized by oval or round conformation and a well-defined and circumscribed shape. Their contour is usually regular, without spicule and lobes. The degree of infiltration into surrounding tissues is low or very-low by virtues of its non-aggressive nature.

On the other hand, malignant tumors usually turn out to be irregular in their profile with ill-defined, star-shaped, and lobulated boundaries. These masses are inclined to damage surrounding tissues, thus they are characterized by a high degree of infiltration which results in inhomogeneous margins and a great number of spicule and concavities [12].

1.2 Current Breast Diagnostic Imaging Modalities

The process of detecting breast cancer usually begins with an examination carried out by the doctor, who checks both breasts (and the lymph nodes next to them) to identify by touch any masses or abnormalities.
Different imaging modalities based on different physical techniques can be used for breast cancer detection and diagnosis. The most common methods are Digital Mammography, Digital Breast Tomosynthesis (DBT), Breast Ultrasound (US) and Breast Magnetic Resonance Imaging (MRI).
The procedures, which most of breast cancer prevention and treatment centers usually provide, include Mammography for screening, DBT, Breast MRI and Breast US for evaluating abnormalities, and the biopsy for laboratory testing suspicious masses identified in the previous steps [13].

In particular, breast biopsy involves the removal of a small breast tissue sample to conduct a histological study and is essential to determine the nature of any possible lesion or suspected area in the breast. However, although the goal is to intervene only if the masses could be cancer, the number of negative biopsies (i.e. low-risk benign tumors), corresponds to approximately 75% of the total [14]. These numbers lead to several downsides both for the patient who perceives discomfort and pain and may have problems at the biopsy site (e.g. bleeding or infection), and for the National Health Services that will have to cover higher healthcare costs. Part of this condition can be attributed to the previously mentioned breast diagnostic imaging modalities, which present both positive aspects and criticalities to achieve the purpose for which they are part of the clinical routine.

1.2.1 Mammography

2D Full Field Digital Mammography (FFDM) is the most commonly used technique for breast cancer detection. Although the technologies available for prevention and diagnosis are numerous, today it represents the *gold standard* of breast cancer screening [15]: the enrolment in these screening programs has the advantage of reducing breast cancer mortality rate by 40% [16].
It is an x-ray imaging technology that consists of projecting a ray beam that passes through the breast compressed between two plastic plates and reaches a detector on the opposite side. From the intensity measurement of these rays, it is possible to distinguish the different tissues present inside the breast and detect possible lesions (cysts, fibroadenomas, calcifications).
The energy of the X-rays used in mammography ranges from 17.5 keV to 22.7 keV, depending on the anode material (typically Molybdenum or Rhodium), and is lower than those used for bone radiography (up to 70 keV if tungsten anode) [17].
Despite the advantages in terms of mortality rate reduction of early detected cancers, several negative aspects should be emphasized.

Indeed, this technique has proven itself unfit for dense breasts, and other complementary tests such as DBT, US, and MRI are necessary, especially for young women who have more glandular tissue (high density) than adipose one [18].
Moreover, breast compression involves problems both for patient comfort and structure overlapping that can hide abnormal masses under dense healthy tissues.
In particular, the pressure that is applied during the examination is an important issue because it may lead to a reduction of sensitivity or specificity if it is too high or too low, respectively [19].

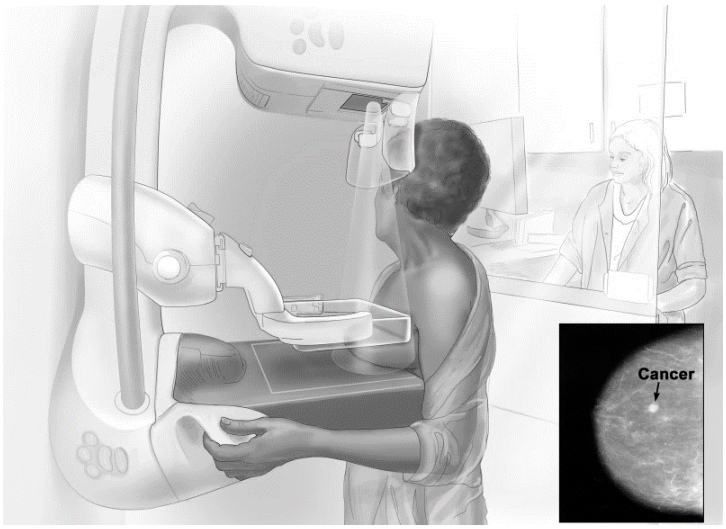


Figure 4. Illustration of a screening mammography exam. The standard exam consists of a bilateral mammogram in two views, which are the craniocaudal (CC) and the mediolateral oblique (MLO). The sketch here reported represents the configuration to obtain the CC projection. Retrieved from [20].

1.2.2 Digital Breast Tomosynthesis (DBT)

Digital Breast Tomosynthesis (DBT), also called 3D Mammography, is the evolution of the digital 2D mammography and is configured as an imaging technique that generates pseudo-three-dimensional images by acquiring several 2D projections of the breast over a limited angular range (typically 10°-50°) [21].
Its acquisition technology and its implementation in clinical practice can partially solve the problems related to the superimposition of overlapping tissues because it produces a pseudo-3D breast reconstruction.

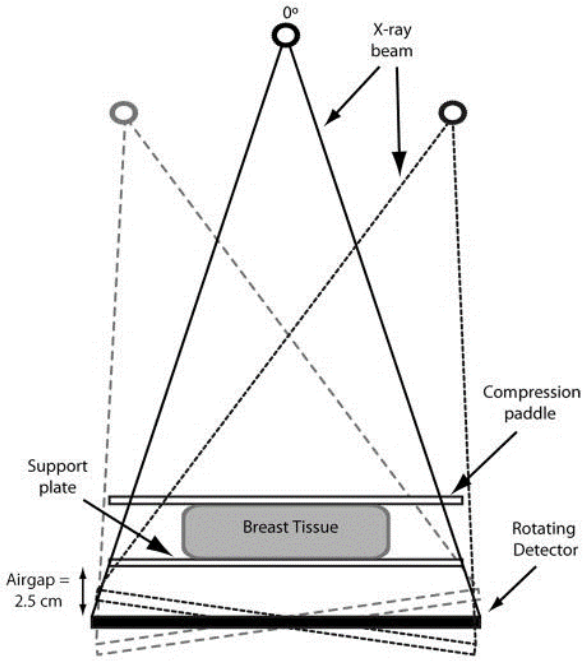


Figure 5. Schematic representation of DBT operating principle. The system acquires a limited number of projections over a certain angular range (e.g. 15 projections over 15°) to get information from multiple points, which however do not cover the whole breast volume. Retrieved from [22].

Clinical studies have demonstrated that DBT increases the sensitivity and specificity currently achieved in mammographic screening tests, especially in the case of dense breasts (40% reduction in the false-positive recall) [23].
Moreover, DBT only provides for slightly higher levels of radiation than mammography because it acquires more but low-dose exposures of the compressed breast. However, it is not yet the standard in breast cancer screening, but there are current clinical studies that are investigating the feasibility of implementing DBT protocols in screening programs [22, 23]. In particular, DBT data would be used to reconstruct not only the pseudo-3D breast volume but also the 2D images used in the screening examination to reduce patient exposure [24].

1.2.3 Breast Ultrasound (US)

Breast US is a simple, safe and non-invasive diagnostic exam since it is based on the emission and reception of low-frequency and high-intensity sound waves, which do not cause damage to the organism. The US emitted by the probe are reflected differently depending on the type of tissue, creating a grayscale image that makes it possible to identify potential abnormalities inside the breast, distinguishing them between those filled with fluid (e.g. cysts) and those that are solid (e.g. fibroadenomas, or cancerous masses).

Breast US is not a stand-alone screening test but an examination that completes the mammography in all those cases where its performances are not adequate and the lesions cannot be detected with a proper level of confidence: indeed, sensitivity and specificity of breast US are higher than mammography in young women and those with dense breasts [25].
Therefore, both imaging modalities are used in most screening programs: for instance, the Food and Drug Administration (FDA) approved the Automated Whole Breast US (AWBUS) as a supplemental tool for screening women with extremely dense breast tissue [26].
Breast US cannot be adopted as the only screening modality due to its high false-positive and false-negative rates, associated with the need for more staff training, additional studies and integration into current PACS [26, 27].

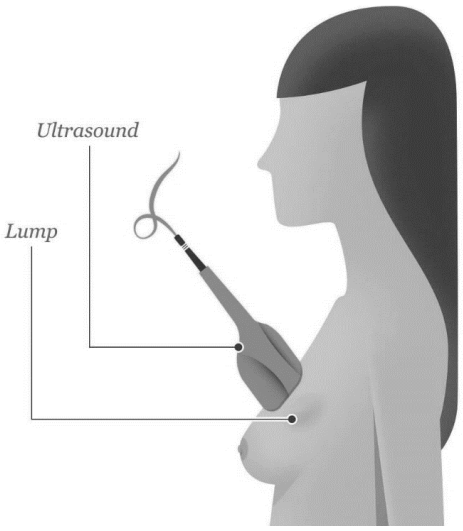


Figure 6. Simple representation of the breast US exam. The image obtained is called sonogram and is useful for assessing whether the mass is solid (e.g. fibroadenoma) or filled with fluid (e.g. cyst). Retrieved from [9].

1.2.4 Breast Magnetic Resonance Imaging (MRI)

Nuclear magnetic resonance (NMR) is a multi-planar diagnostic technique that provides detailed images of organs and body structures using magnetic fields and radio waves, without exposing the patient to any type of ionizing radiation. In particular, the patient is immersed in a high static magnetic field (0.5-3 T), which orients the axis of H+ protons in the patient body fluids along the field itself. Then, radio waves with an appropriate frequency (i.e. Larmor frequency) are generated to make protons resonate and provide them with energy. Due to temporary deformations of the nuclei, the H+ atoms emit signals, which differ according to the composition of each tissue where they are, and are detected by coils positioned on the breast.

In the case of breast MRI, it is necessary the use of contrast agents to temporarily change the tissue properties and discriminate structures with similar magnetic behavior: this imaging modality is known as Contrast-Enhanced Magnetic Resonance Imaging (CE-MRI).
The images obtained from the recorded signals are characterized by more attention to the soft tissues, and, in the case of breast imaging, this allows them to reach very high sensitivity values (more than 90% if CE-MRI) even with dense breast [28]. Instead, its low-moderate specificity (which ranges from 50% to 80%) results in problems for lesion characterization and discrimination, which, therefore, would cause a high number of unnecessary breast biopsies and follow-ups if it replaced mammography as a screening tool [28, 29]. For this reason, breast MRI is usually performed after a positive biopsy to obtain more information about cancer and its body spread. It is also used for screening with mammography for those women at high risk of cancer contraction (e.g. genetic mutations, family history of breast cancer) [30].
Several breast MRI limitations can be mentioned: equipment complexity, high cost (approximately ten times of mammography), time-consuming exam, non-negligible toxic component of the currently used intravenous contrast medium (i.e. Gadolinium) [31]. Nevertheless, it is configured as an imaging modality that has shown its value in breast cancer assessment both in terms of detection and staging.

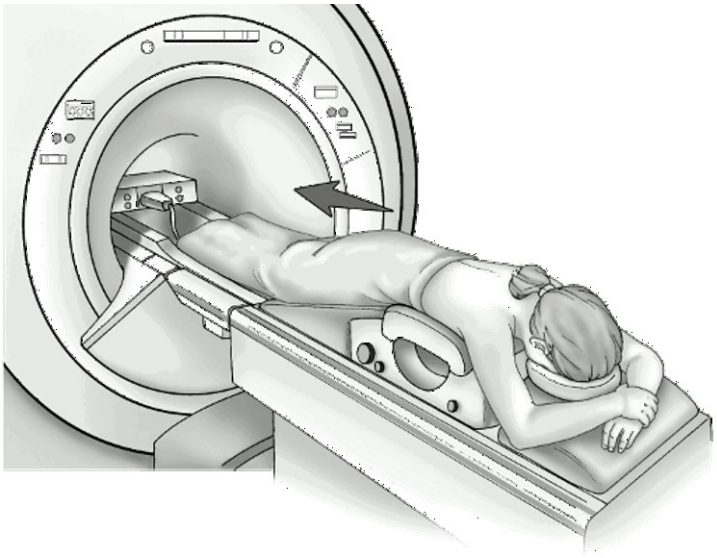


Figure 7. Drawing representing the execution of a breast MRI. Before performing this diagnostic exam, a contrast medium (typically Gadolinium) is injected intravenously to help differentiate the structure components of the breast. Retrieved from [32].

1.3 Dedicated Breast Computed Tomography (DBCT)

Dedicated Breast Computed Tomography (DBCT), also known as 3D Breast CT, is a new breast imaging modality (the first commercial prototype was available from 2006 [33]), which aims to improve breast cancer detection and diagnosis.
Indeed, DBCT is not yet implemented in daily clinical practice, but the numerous researches carried out in the centers that own this system (3 in the United States, 3 in China, 1 in Qatar, 1 in Thailand, 1 in The Netherlands) want to demonstrate the high technological level of this imaging method, which allows overcoming the main criticalities of other conventional breast diagnostic imaging techniques.

DBCT is a computed tomographic system that provides a true high-quality 3D morphological image of the breast. The operating mechanism is based on the collection of several x-ray beams taken from different angles (*angular sampling*) to reconstruct the spatial distribution of the linear coefficient μ. Starting from this distribution map, the 3D visualization of the breast internal structures is reconstructed retrospectively, typically using a *Filtered Backpropagation* algorithm. Differently from DBT, DBCT gantry realizes a relative fast full rotation around the pendant breast on a horizontal plane (Figure 8), and its x-ray source and detector acquire 30 projection images per second (typically 300 frames over 10 seconds of acquisition sequence) [34].

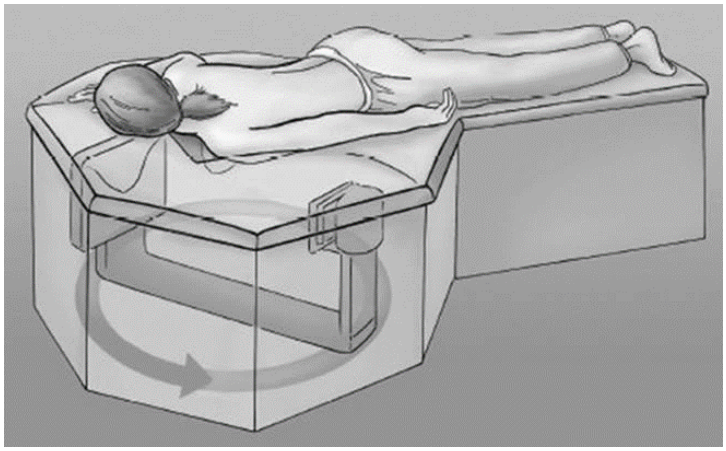


Figure 8. Representative illustration of the DBCT structure. Both the x-ray source and detector motion plan and the patient's position can be observed. The patient is lying prone on the table with the scanned breast inside the hole in a pending configuration. Retrieved from [35].

Many of the DBCT advantages can be found in the new acquisition method and the breast geometric configuration. From a technical point of view, DBCT is the first true fully 3D breast imaging technique that enables uniquely and decisively to overcome the tissue superimposition issue that occurs during mammography (and partially DBT) [36, 37]. Moreover, the breast is not compressed because it is positioned inside the table aperture (like Breast MRI), favoring both image quality and patient

comfort. On the one hand, the complete 3D reconstruction from the chest to the nipple allows a detailed lesion visualization in all types of breast (size and density) [38], on the other hand the examination does not ask the patient for breast compression, and, compared to mammography and DBT, is significantly more comfortable and less disturbing [35].

Another advantage not to be underestimated is its spatial resolution because its isotropic voxel size can go below 300 μm [34, 39]: DBCT identifies itself as the highest definition tomographic technique when compared with conventional breast imaging methods as DBT (typically 0.1 mm x 0.1 mm x 1 mm [40]) or breast MRI (typically 0.7 mm x 0.7 mm x 1.3 mm [41]). In the near future, DBCT can find its space for its use in the clinical routine by the greater resolution that will allow obtaining a greater detection and diagnosis accuracy than that of the imaging techniques currently adopted. Furthermore, thanks to the dedicated application to breast, x-ray tube and additional filtrations are designed to meet the specific contrast requirements of breast tissues, thus maximizing the resulting image contrast.

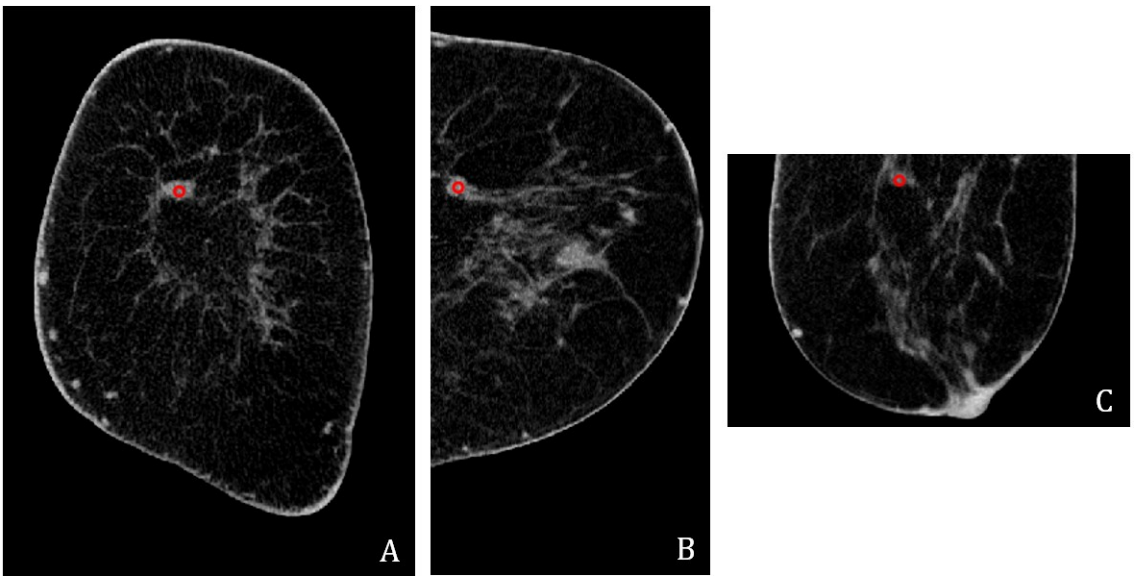


Figure 9. Coronal (A), sagittal (B) and axial (C) DBCT images of the same breast that visualize a cyst (marked by the red circle). It presents the typical distinctive features of a benign tumor by its round shape and well-defined margins. These projections show the high spatial detail that can be obtained with this imaging modality: it can improve the discrimination of the different breast lesions. They are part of the dataset used in this study. Figure courtesy of Marco Caballo and Ioannis Sechopoulos.

As regards the radiation dose level, it has been shown that it is possible to obtain an average glandular dose (AGD) comparable to that of a conventional 2-view screening mammography, even if its level increases if adopted for diagnosis or women with large or dense breast (but always below the diagnostic mammography levels) [42, 43].

Moreover, DBCT radiation dose distribution is more uniform than mammography [44, 45], and its quantity can be decreased with an optimization of the acquisition process, both hardware (i.e. detectors) and software (i.e. imaging reconstruction).

Obviously, emphasizing its role as a new breast imaging modality and considering the numerous real projections (between 300 and 500 per scanned breast), the amount of data available to radiologists requires not only a longer reading and interpretation time than that of the other techniques (primarily, mammography and DBT) but also a training to understand the importance of visualizing and evaluating the different lesions in their true high resolution 3D nature. This aspect can raise questions about the actual adoption of DBCT in the breast cancer assessment process (especially in the screening routine), but surely makes this imaging technique an additional useful diagnostic tool to be used in parallel or in place of DBT, breast US, and breast MRI. For this purpose, Computer-Aided Detection (CADe) and Computer-Aided Diagnosis (CADx) systems could be beneficial because they will exploit DBCT to its full potential, and support the radiologists in reading these new medical images, and making better and more complete clinical decisions.
In particular, the extraction of quantitative biomarkers from these radiographic images, also called *Radiomics* (which will be discussed in Chapter 2), provides an opportunity for breast tumor recognition and assessment, especially considering the high spatial and contrast resolution of DBCT images. The combination of the radiomics approach with the modern and advanced machine learning (ML) techniques applied also in digital radiology constitutes an important breakthrough-point for the detection and diagnosis of breast mass-like lesions.

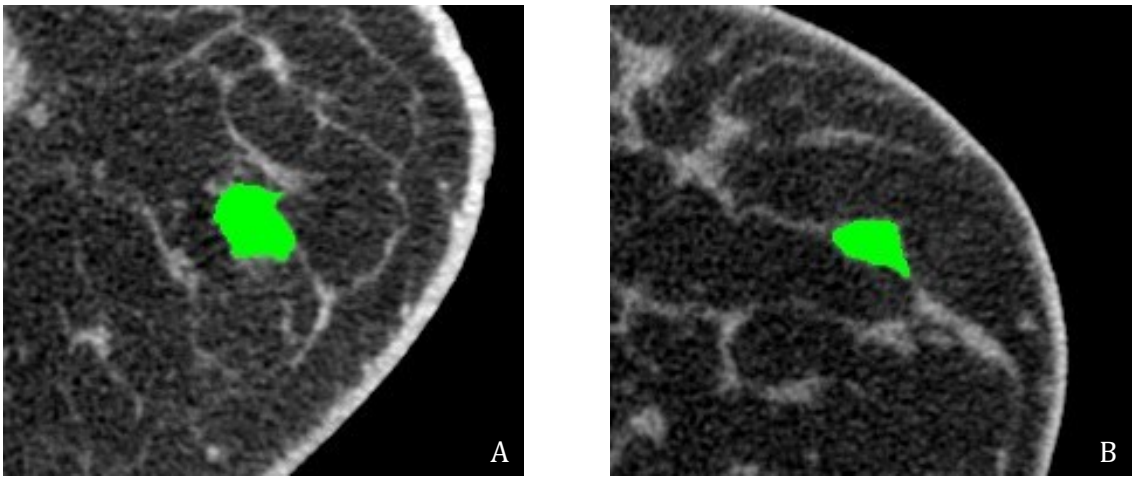


Figure 10. Two coronal DBCT patches of the same breast containing two different cysts (benign tumors) manually segmented. These patches are part of the dataset used in this study. Figure courtesy of Marco Caballo and Ioannis Sechopoulos.

1.4 Thesis Content

1.4.1 Objective of the Study and Motivation

The goal of this thesis is to design a Computer-Aided Diagnosis (CADx) system for breast mass-like lesion classification in Dedicated Breast CT (DBCT) images using a quantitative radiomics approach and a multi-layer perceptron artificial neural network (ANN).

As concerns the breast tumor characterization, this work was focused on the design, development, validation, and implementation of new descriptors that quantify the tumor properties in terms of morphology (i.e. shape and contour) and margin appearance (i.e. sharpness and infiltration degree). This choice was linked to the interest of evaluating all those features that, in most research studies, are not taken into consideration (at least not to a large extent) since many of the previously published studies only include the well-known and traditional texture features.

With regard to the classification task, this work involved a study related to the ANN architecture in terms of number of hidden layers and hidden neurons (*Neural Network tuning*), the number of features to consider (*feature selection*), and the number of samples to be used during the network training (*dataset balancing*) in order to correctly discriminate benign and malignant masses.

The clinical motivation of this work is to reduce the number of negative and unnecessary breast biopsies, which constitutes more than 70% of the overall biopsies performed [Section 1.2]. Therefore, the purpose is to develop a radiomic-based classification algorithm operating on DBCT images, which provides help to make decisions about breast tumor diagnosis.

1.4.2 Thesis Outline

This thesis describes the work carried out at the Advanced X-ray Tomographic Imaging (AXTI) Laboratory, Department of Radiology and Nuclear Medicine, Radboud University Medical Center (Nijmegen, The Netherlands), during the period March 2019 - September 2019, and is divided into the following five chapters.

Chapter 1 (*Introduction*) outlines the context in which the project was defined. In particular, breast cancer, the current imaging breast diagnostic imaging methods used, the DBCT imaging modality adopted in this work, and the purpose and motivations behind it.

Chapter 2 (*Radiomics*) illustrates an in-depth study on the radiomics approach, introducing its definition and peculiarities, and showing what has been achieved so far in the previous research studies applied to oncological imaging with a particular focus on breast cancer.

Chapter 3 (*Materials and Methods*) shows the proposed radiomic pipeline, the image dataset available for this study, the developed and validated features, and the adopted machine learning classification model.

Chapter 4 (*Results*) reports the results obtained at the end of the training and the evaluation processes for selecting the best ANN classification model based on the developed shape and margin radiomic descriptors.

Chapter 5 (*Conclusions*) delineates the final observations on the potential outcomes arising from the developed system, the possible future works that could be performed, and the main constraints of this study.

2 Radiomics

2.1 Definition and Workflow

Innovation in the field of diagnostic imaging does not only happen in the technological evolution of imaging equipment and the optimization of acquisition procedures, but also the development of medical image analysis methods.
This is the context of *Radiomics*, a discipline that aims to extract mineable information from medical images. Indeed, medical images are nowadays no longer considered only as visually perceived information, but constitute a source of data that can be extracted and quantitatively analyzed to assist physicians and radiologists in their clinical decisions [46]. Radiomics assumes that a quantitative analysis of the target region (i.e. the region of the image with prognostic value) through a set of texture and morphological features can provide useful data to guide the clinician not only towards a correct detection and diagnosis, but also, if the case, towards predictive factors for response to treatment. Therefore, these radiomic data can be extracted from images, and subsequently used for the development of mathematical models that may potentially improve all areas of research and clinical application related to oncological imaging.
Radiomics can be seen as an evolution of the traditional CADe and CADx systems (which date back to the 1980s), but has two innovative aspects [47]:

- it considers a very high number of features (hundreds or even thousands), while CAD systems typically include less than a dozen features;
- its investigation aims at focusing on many cancer-related applications, such as detection, diagnosis, survival prediction, cancer staging, often combining image-based information with patient data and genomic expression (in this latter case, it is commonly referred to as *radiogenomics*). As opposed to this, CAD systems are usually used only for detection and diagnosis.

The radiomics workflow can be articulated in four main steps [48], which are described as follows:

1. Acquisition and reconstruction of high-quality biomedical images obtained through the different imaging modalities (usually 3D imaging, such as MRI or CT).

2. Segmentation of the prognostic region (e.g. the tumor) by an experienced radiologist or by automatic or semi-automatic software.

3. Feature extraction starting from the segmented regions of interest (ROIs). Different categories of features used within the radiomics context refer to the pixel intensities, the texture patterns, the geometric shape, the contour

irregularities, and the interaction of the lesion with the surrounding tissues (e.g. infiltration).

4. Development of the predictive (or prognostic) model starting from the features considered useful and influential for the final goal (e.g. diagnosis, survival prediction) through a *feature selection* process that removes all those redundant and non-relevant variables.

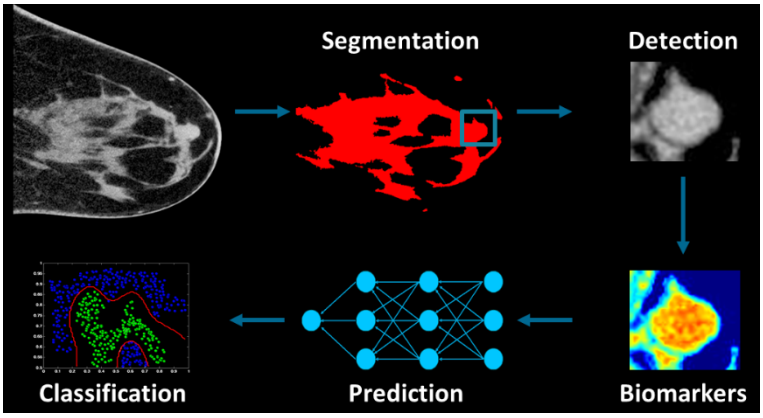


Figure 11. Example of a quantitative radiomics pipeline. In this case, the breast is the target organ and the images acquired are dedicated breast CT (DBCT). Here, the final goal is breast tumor diagnosis, i.e. classifying the masses as benign or malignant. Retrieved from [49].

Going more into detail about the *feature extraction* step, we can identify some types of features, which have different names depending on the state-of-the-art studies taken into consideration.
By referring to the work of Kumar et al. [50], the three main types of radiomic features are those based on intensity, texture, and shape.
Intensity-based features, traditionally called *first-order statistics*, are the simplest to implement. Indeed, they describe the ROI in statistical terms, extracting the most common first-order statistical descriptors from the set of intensity values of the pixels belonging to the segmented region. Among them, mean, standard deviation (STD), median, maximum, minimum, entropy can be listed. In some studies, intensity-based features are considered as closely related to texture-based ones, which are traditionally used in most of the radiomic pipelines found in the literature.
Texture-based features, also known as *second-order statistics*, statistically describe the spatial relationships between the various pixels within the ROI. Several groups of features can be included in this category [39], and are briefly described below:

- The aforementioned **first-order statistics features**, which are based on the intensity values distribution of individual pixels without reference to their mutual relationships.

- **Haralick features** [51], which describe the inter-relationships between neighboring pixels. In particular, they are extracted starting from the Grey-Level Co-occurrence Matrix (GLCM), which computes the number of occurrences of pixel pairs with specific values and specific spatial relationships (Figure 12). More GLCMs at different angles (e.g. 0°, 45°, 90°, 135°) are usually calculated to consider a higher number of adjacency spatial relationships. Several descriptors (e.g. contrast, energy, correlation) are extracted from each of these matrixes, and, finally, their average returns the single final values.

- **Run-length features** (or Galloway features) [52], which assess the homogeneity of the ROI for each grey-level. Similarly to Haralick descriptors, they are extracted from an appropriate matrix called Grey-Level Run Length Matrix (GLRLM), which considers the length of consecutive pixels having the same intensity along a certain angle (Figure 12). Different descriptors, such as run percentage and grey-level and run-length non-uniformity, are then calculated from these representations.

- **Structural and Pattern features** [53], which analyze the composition of the ROI architecture and the local intensity variations. Hessian, Fractal, and Laws features are part of this texture-based category.

- **Gabor features** [54], which provide an analysis of the frequency content within the ROI in specific and localized portions. Two-dimensional Gaussian kernels modulated by a sinusoidal plane wave are convoluted with the input image along many different orientations at different frequencies. Finally, the mean and standard deviation are calculated to have single values for each ROI.

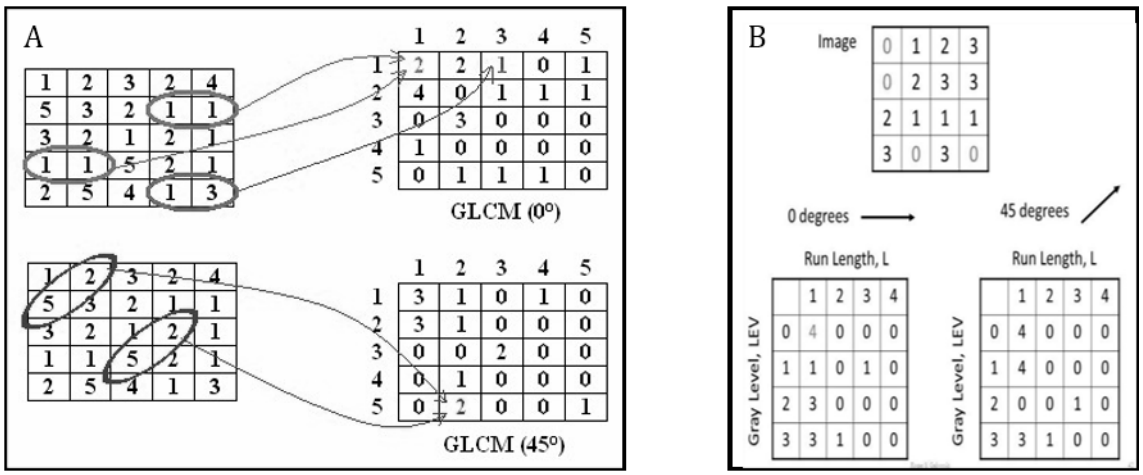


Figure 12. Simple examples of the construction process of a GLCM (A) and of a GLRLM (B), from which Haralick and Galloway features are extracted, respectively. Retrieved from [55] and [56].

Instead, *shape features* are based on the segmentation mask of the lesion and try to quantify region morphology and contour irregularities. These are usually incorporated in radiomic pipelines, but often refer to simple measurements (e.g. area, perimeter, eccentricity) [57]. Moreover, the inclusion of these descriptors in previously proposed studies is lower than intensity-based and texture-based features. In particular, some investigators came up with the development of few novel descriptors, but, only in some works, these newly developed features founded an application in diagnostic pipelines. That is why designing additional, more complex, robust shape features to include in radiomic pipelines, holds considerable relevance. It will be further investigated and described in Chapter 3 (*Materials and Methods*) since they constitute an essential part of this thesis.

For the sake of clarity, the radiomics approach can be applied directly to 3D images obtained from tomographic imaging techniques. Therefore, it is possible to extract features that directly quantify the characteristics in this dimensionality, but the implementation difficulties and the high computational cost required in the case of many 3D radiomic descriptors have led to the prevalent use of 2D radiomic features in the vast majority of works. To investigate similarities and differences, studies have been carried out to compare the results achieved with both methods, discovering that the 2D radiomics performance is comparable or better than in 3D (Figure 13), with the advantage of a simpler mathematical formulation [58].

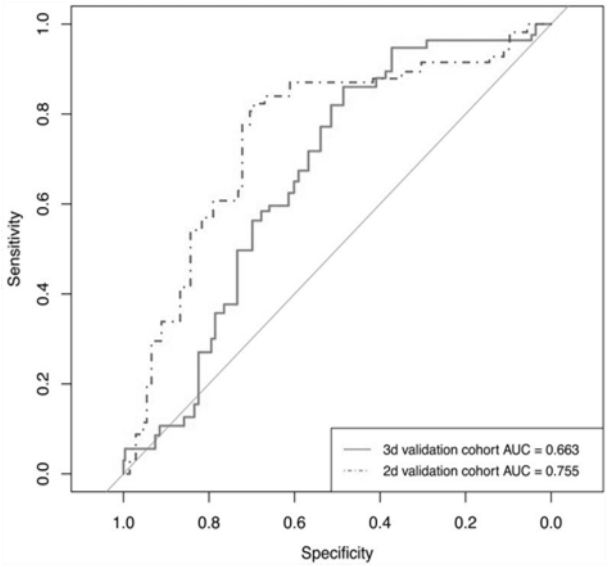


Figure 13. Receiving Operating Characteristic (ROC) curves obtained from the application of 2D and 3D radiomic descriptors on a validation set of CT images. These curves, usually created by plotting the sensitivity against the specificity (or parameters derived from them), constitute a useful tool for evaluating the performance of a binary classifier. In this case, they show how the 2D radiomic model behaves slightly better than the 3D one. Therefore, 2D features are preferable, even considering the lower development effort and the lower computational cost. Retrieved from [58].

2.2 Radiomics in Oncological Imaging

Radiomics has found its application in all areas where medical imaging plays a fundamental role in the detection, diagnosis, and prognosis of cancer. The main use and successes have occurred in the studies related to the most frequently diagnosed cancers: in particular, radiomics has achieved satisfactory performance in lung, breast, prostate and liver imaging.
Putting aside the breast cancer state-of-the-art that will be shown in Section 2.3, some interesting radiomics studies focused on the other cancers that have been a starting point for this thesis approach are described below.

Zhang et al. [59] developed a prediction model of non-small lung cancer cell recurrence based on lung CT images. They included several radiomic descriptors, which can be subdivided in intensity and texture features, then applied different methods of feature reduction. This process is useful not only to eliminate descriptors that are simply noise or are redundant compared to more highly predictive variables but also to avoid overfitting when machine learning (described in Chapter 3) is used for outcome prediction based on the extracted and selected radiomic features. Indeed, overfitting occurs more easily when the number of features is much higher than the available dataset size, as it often happens in medical studies. Furthermore, they also analyzed different ML models, including Random Forest (RF), Support Vector Machine (SVM), and Neural Network (NN). The AUC values were adopted to compare the different combinations obtainable from the feature selection FS and classification methods: the RF combined with a previous Principal Component Analysis (PCA) feature selection and a dataset balancing with oversampling of the minority class led to an AUC of 0.79, a value usually considered satisfactory for cancer recurrence prediction problems.
Aerts et al. [60] conducted the first extensive clinical radiomics implementation model for survival prediction in the case of lung and head-neck cancer based on CT images of 1019 patients. Their initial features were 440 and belonged to the three categories abovementioned (i.e. intensity, texture, shape). They studied the stability and prognostic performances of all features, obtaining promising results that led to a deeper analysis of the clinical impact of this approach (Figure 14).
Griethuysen et al. [61] carried out a study always based on lung CT images, which led to the development of a classification model of lung nodules. Their dataset consisted of 429 distinct lesions, and their features set included 1120 radiomic features, belonging to the groups of first-order, textures, shape, and higher-order (e.g. wavelet) features. After performing a stability study to preserve only the most stable, the classification results showed significant differences between the group of benign and malignant lesions.

Vignati et al. [62] carried out a study to evaluate the possibility of a radiomics approach applied to T2-weighted and diffusion-weighted MRI images for the evaluation of prostate cancer aggressiveness (Gleason score). Their interest was to demonstrate that texture features could help in defining the cancer stage and, based on this, the following tailored treatment. The adoption of these two imaging methods on 45 patients made it possible to reach AUC values above 0.92.
Wibmer et al. [63] performed a study of 147 patients to show the potential for detection and cancer grading of the Haralick features on T2-weighted and diffusion-weighted MRI images. Their work analyzed the associations between the texture-based descriptors and the cancer Gleason score and concluded that these features could be useful for adding more information in the decision-making process about the cancer progression.

Finally, Zhou et al. [64] developed a CT-based live cancer recurrence prediction model. The work was based on a retrospective analysis of 215 patients, and 21 texture-based features were implemented. In particular, they showed that the addition of radiomics to the current clinical model provides superior performance in the early recurrence discrimination with a sensitivity of 82% and a specificity of 70%. Therefore, radiomics demonstrated to be an approach with high prognostic abilities and of great utility in opting for the best treatment.

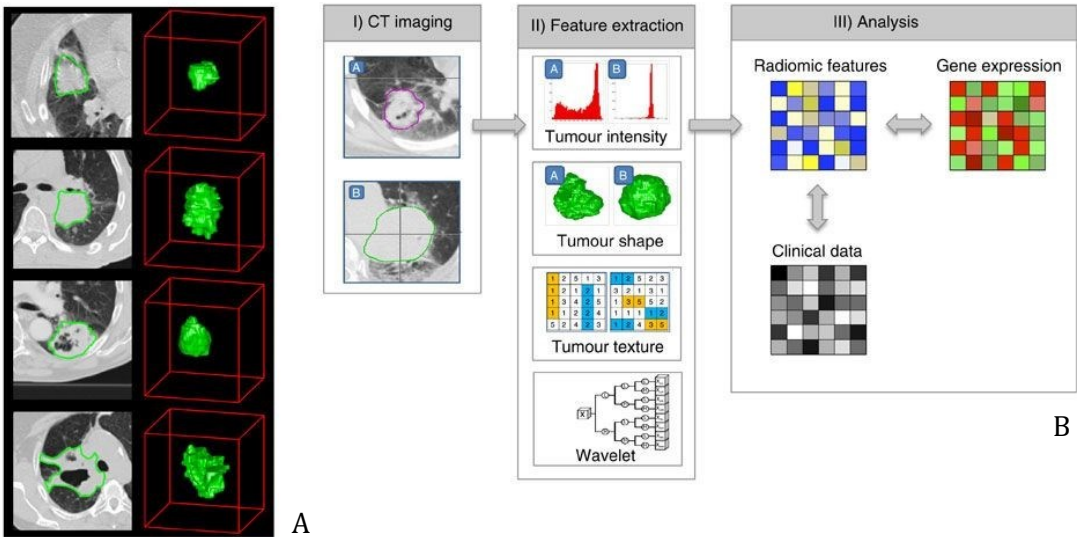


Figure 14. Example of lung cancer CT images (A). Representation of a radiomics pipeline that starts from image acquisition and tumor segmentation, then followed by feature extraction and data analysis (B). Retrieved from [60].

2.3 Radiomics in Breast Cancer Imaging

One of the main fields where radiomics is achieving growing success and where progress in terms of image processing is tangible is breast cancer imaging.
In literature, it is possible to find studies that adopt all the different imaging methods described so far. The project goals are more directed towards the detection and diagnosis of tumors, but there are also works where the possibility of providing support to doctors in defining treatment strategies or understanding the role of the radiomic approach in the prediction of recurrence-free survival is evaluated.

Many of the studies are based on digital mammography images since mammography is the preferred exam for screening and also serves in the subsequent phase of diagnosis, which makes it possible to have many available images.
Mao et al. [65] carried out a study that involved 173 patients and aimed at defining the ML algorithm that could correctly classify benign and malignant masses. Four different algorithms were built, including SVM, Logistic Regression (LR), K-Nearest Neighbor (KNN), and Bayes Classification. At the end of the feature selection process, each ML model contained a total of 51 radiomic features (intensity, texture, shape). The LR classification model showed the best results in terms of accuracy, sensitivity, and specificity (0.89, 0.87, and 0.9, respectively), higher than those achieved by radiologists.
Another interesting work is that of Li et al. [66], which compared the current clinical flow and a radiomics approach for the prediction of invasive carcinoma from DCIS lesions, starting from mammography images acquired from 250 patients. All types of features were included in the initial feature set, but texture-based and morphology features were those mainly extracted (Figure 15). Five different feature selection approaches and seven types of classifiers were combined, and the final statistical analysis showed how the radiomic descriptors could complement the current clinical characteristics to identify invasive cancer starting from DCIS. Another point of this study to be underlined is that the features most present at the end of the feature selection process and those that provided for greater predictive value were the morphological radiomic ones.

Chan et al. [67] used DBT images for the characterization of tumors, making a comparison between different ML methodologies applied both to the acquired projective views and to the DBT reconstructed slices. The peculiarity of the feature set was the strong presence of features that analyze the spiculation of the lesions. The Linear Discriminant Analysis (LDA) classifier with a stepwise feature selection was adopted. This study showed that the DBT slice-based approach had higher performance since the AUC ranged from 0.87 to 0.93, against a range from 0.78 to 0.84 of the acquired projections data. Therefore, the authors concluded that it is worth building diagnostic models starting from the DBT pseudo-3D volumes.

Moreover, it is possible to identify some works that have used US images.
In particular, Lee et al. [68] developed a radiomics model that allows diagnosing starting from the application of texture features in US imaging. Their dataset included 901 lesions between fibroadenomas and triple-negative breast cancer (TNBC). The final AUC value was 0.84 on the validation set, a satisfactory outcome considering that these two types of tumors are difficult to distinguish via simplistic visual perception.
Nugroho et al. [69] also constructed a classifier for breast nodules diagnosis based on the adoption of only features related to the margins of the lesions. Indeed, the interesting aspect of this study was to consider only features (even developed from scratch) that would analyze only the outermost portions and the relationship with the surrounding environment of the breast lesions. By implementing a NN, they achieved an accuracy of 0.95, a sensitivity of 0.93, a specificity of 0.96, and an AUC of 0.99, which led them to the conclusion that this margin-based approach could certainly be valuable for aiding the radiologists.

As regards MRI imaging, Huang et al. [70] focused on the prediction of breast cancer recurrence-free survival using radiomic descriptors applied to 113 patient PET and MRI images. The features were mainly texture-based, and several ML algorithms were tested. The results enabled them to assert the usefulness of their models, reaching a maximum of 0.75 in the AUC values. Considering that the survival prediction is a remarkably difficult task, an AUC value above 0.7 is often judged quite good because not even a radiologist can say anything about the cancer stage by only looking at the images.
Xiong et al. [71] wanted to investigate the issue of predicting breast cancers that are insensitive to certain drugs, so it concerned the cancer treatment area. This study involved a total of 125 patients who did an MRI exam before undergoing neoadjuvant chemotherapy. More than 1900 variables were previously extracted, only to be reduced by feature selection. An LR classifier was constructed by combining selected descriptors and clinical risk factors: this combined model showed higher performance than the simpler clinical one. In particular, it reached 0.93 of accuracy and 0.93 of AUC against 0.87 and 0.79, respectively.

Finally, a brief analysis of all those works that adopted the same imaging technique as this thesis can be provided. Most of the DBCT-based studies include detection and diagnosis applications.
In particular, Reiser et al. [72] developed an automatic detection method to include in a CADe. They had 132 cases in their dataset, balanced between benign and malign masses. Unlike most of the works, they implemented a series of 3D shape-based and margin-based features and adopted an LDA classifier. They reached a sensitivity of 0.84, but they also evaluated the algorithm performance in the case of different breast densities: as expected, results get worse if the breast has a greater density,

given the high masking effect of glandular tissue on breast masses (AUC of 0.76 compared to 0.86 for the lower density group).
Ray et al. [73] designed a CADx that can automatically segment mass-like lesions and implements a NN for the classification task. The model input consisted of the data obtained from the extraction of six texture-based and eight morphology features. They created different NNs that used all the features or only the morphological ones or only the texture ones. They found that the AUC values were 0.80, 0.74, and 0.64, corresponding with all, separated morphological and separated texture-based feature sets, respectively. It is a further demonstration of how useful biomarkers concerning the shape and contour of the lesions could be for breast cancer diagnosis. The results achieved in these two DBCT studies are representative of how an imaging technique with less than ten years of research and development has achieved highly performing results, which are comparable with those obtained with other well-consolidated imaging modalities, despite the lower number of DBCT images available. This suggests that the full 3D nature of DBCT, together with the high resolution and contrast capabilities, may help achieve further insights in breast cancer characterization, especially when combined with state-of-the-art radiomic-based image analysis algorithms.

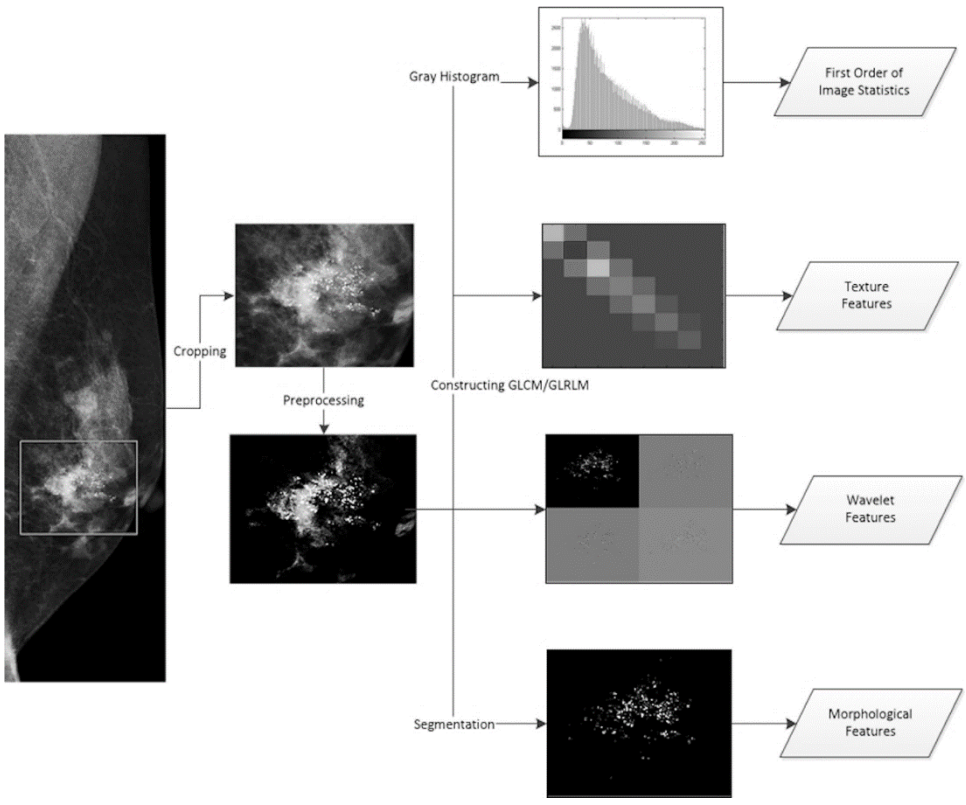


Figure 15. Representation of the workflow followed by Li et al. to extract the features from mammography images. They adopted different types of radiomic descriptors. In particular, there are the first-order statistics features, which refer to the gray-level intensity distribution, the texture-based features, calculated starting from the GLCM and GLRLM matrixes, the morphological features, extracted from the segmentation mask of the lesion, and the higher-order features, whose description is not part of this thesis. Retrieved from [66].

From a comprehensive analysis of the various breast cancer studies analyzed, it is clear that most of them have implemented texture features, and only a small group has chosen to develop and implement features related to morphology and margins. Even in this case, most of the previous works reported the use of a limited number of either shape, or margin descriptors, usually not investigating their combined use in a wider clinical context, nor expanding the implementation to multiple descriptors. Therefore, the goal of this thesis work is to give more insights about shape and margin radiomic descriptors, developing novel features, and evaluating their diagnostic power on a clinical image dataset acquired with DBCT. The intention is to perform a study that not only describes the technical and mathematical aspects of these features, but also investigates their diagnostic capabilities on a real patient-basis, to understand whether these descriptors can help ameliorate diagnostic imaging protocols, in addition to the already well-consolidated textural information, or patient-related data.
Besides, the combination of a new imaging modality (i.e. DBCT) and a radiomics approach that includes multiple types of features could lead to the design of a CADx system with even higher diagnostic potential to support radiologists in their decisions, and to eventually help reduce the number of unnecessary biopsies that many patients with benign tumors nowadays undergo.

3 Materials and Methods

The dataset used and the methods developed and implemented in this study are explained in this chapter. The project pipeline with the main operational steps is shown in Figure 16. All the six macroblocks that constitute the radiomic pipeline will be exhaustively described below.

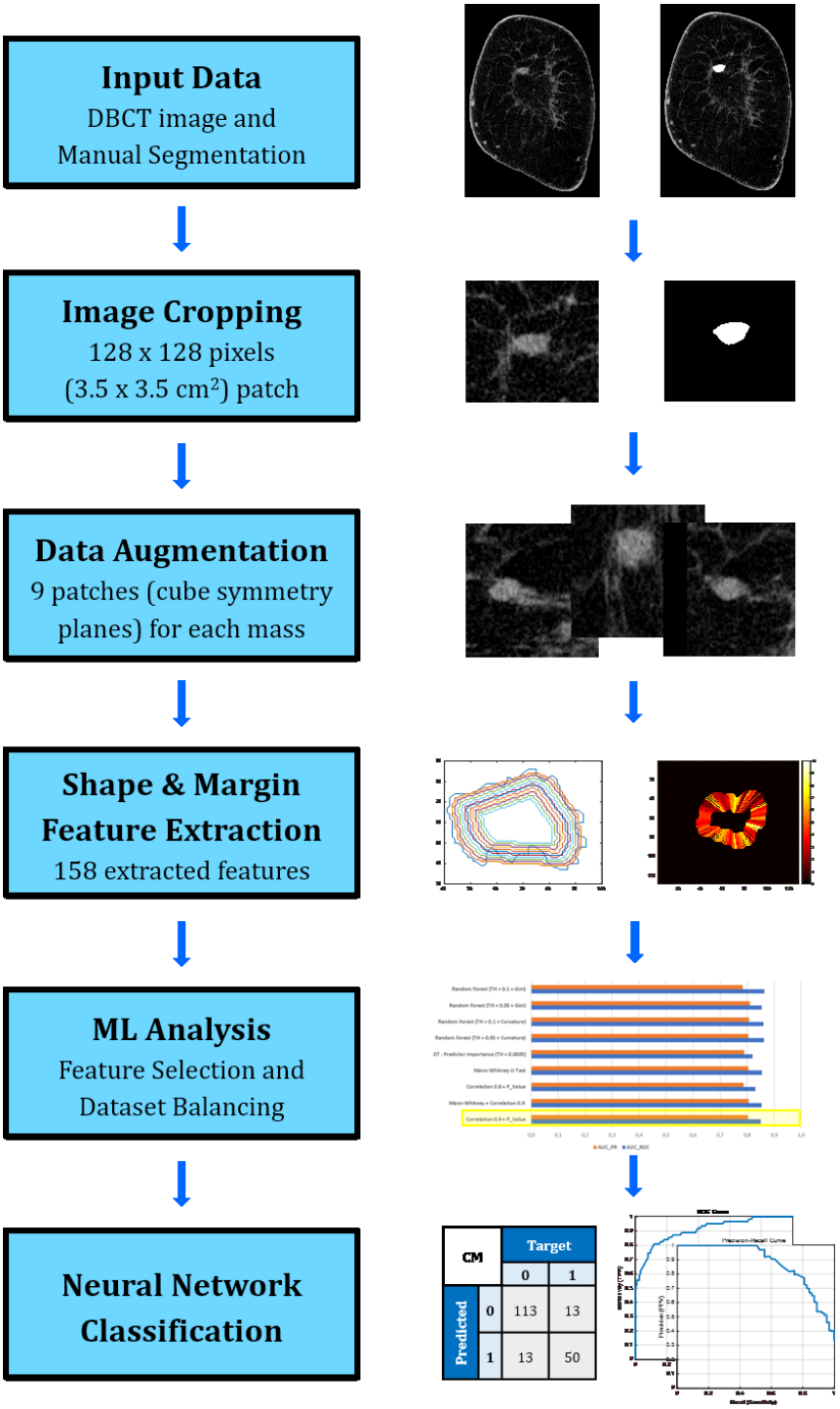


Figure 16. Pipeline of the project. All the operational phases performed are described on the left side, while representative images of each step are illustrated on the right side.

3.1 Image Collection and Annotation

The images used in this study were acquired with the new DBCT imaging modality. The system installed at Radboud University Medical Center is a technology of the *Koning Corporation* (West Henrietta, NY), a manufacturing company that sells only DBCT systems [44, 45]. It is necessary to specify the manufacturer because their device has different technical characteristics than those of other DBCT systems present in other research centers. This system has an x-ray tube (tungsten target, aluminum filter) set to a voltage of 49 kV for all scans with a current ranging from 12 mA to 100 mA according to patient breast size and composition. The average glandular dose for a medium composition and size breast is 8.5 mGy [44]. Regarding the spatial resolution of the DBCT images, the nominal pixel size of the detector is 0.194 mm, while the size of the reconstructed voxel after the application of a Filtered Backpropagation (FBP) algorithm projection is 0.273 mm^3 and is isotropic, differently from what happens for DBT and MRI. For each breast, an average of 300 projections is acquired for each complete DBCT scan, obtained from a whole 360° rotation of the source and detector around the patient breast for a total duration of 10 seconds. All images were taken by radiographers appropriately instructed about the functioning of this new breast diagnosis imaging methodology, while an experienced breast radiologist with in-depth knowledge of DBCT and all breast imaging techniques identified the lesions.

The dataset used in this study consisted of 74 breast masses, obtained from 57 patients over 50 years of age, in which suspected lesions were detected during the mammography screening examination, and who accepted by written informed consent to undergo this additional imaging modality and make their images available for research purposes. These lesions were 20 malignant and 54 benign.
In particular, among the malignant ones, 7 IDC, 7 DCIS, 1 adenocarcinoma, and 5 combinations of types of cancer are identified, while, among the benign ones, there were 46 cysts, 4 fibroadenomas, 3 lymph nodes, and 1 atypical papilloma.
The diagnosis of these masses was obtained by US examination concerning the 46 cysts, while a biopsy was performed for all solid masses to determine their malignant or benign nature. For each lesion, manual annotation was performed by an image analysis scientist under the supervision of a radiologist expert in DBCT, and both raw and annotated mass images were made available for this study. All the original DBCT files were in TIFF format and contained the acquired breast volume as a stack of 2D coronal slices, from the chest to the nipple. Accordingly, the segmentation of each suspicious lesion was performed in the coronal plane, considering all the slices in which part of the mass being analyzed was present. Therefore, for each candidate, two volumes were available, one relating to the original output image at the end of the FBP reconstruction, and one relating to the manual annotation process, which presented the binary mask with only the pixels belonging to the identified lesion.

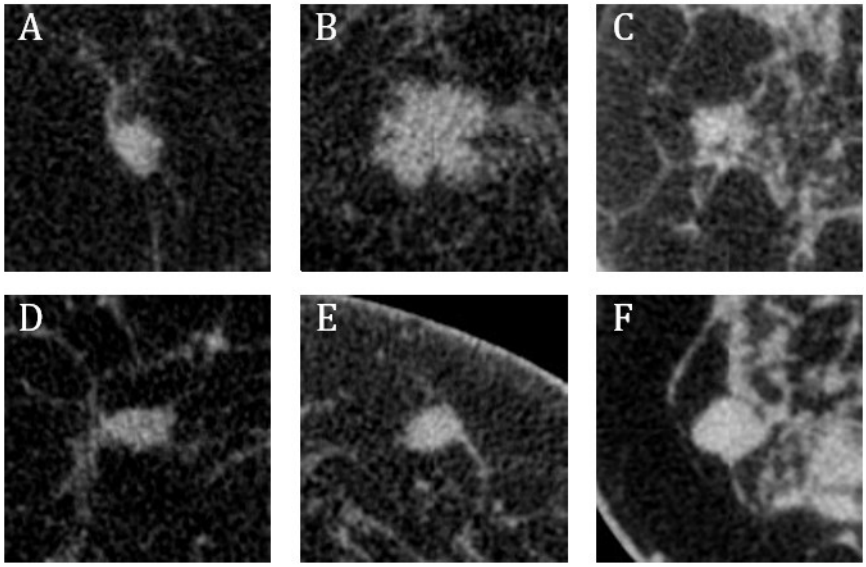


Figure 17. Examples of six different breast lesions belonging to the original dataset. These image patches were extracted from the DBCT coronal slices, and each contained a different mass. In particular, the first three (A, B, C) are malignant and show the typical traits with an irregular shape, a spiculated contour, and a blurred margin. Instead, the last three (D, E, F) represent benign cysts and are characterized by regular shape, well-defined contours, and sharp margins. Figure courtesy of Marco Caballo and Ioannis Sechopoulos.

In addition to the original dataset of real breast masses, a group of phantoms was generated to show the results of the analysis carried out on the various developed shape and margin descriptors, which will be described in Section 3.3.
In particular, the aim was to closely simulate the remarkable features of the benign and malignant tumors to build a solid and relevant phantom study (Figure 18). Therefore, they were both binary masks and grey-scale figures to be able to evaluate both the shape and margin features, which in real lesions analyze the segmentation mask and the grey-scale peritumoral regions of the mass, respectively.

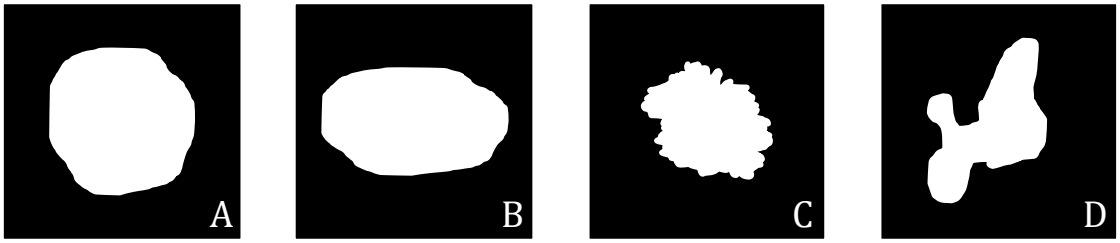


Figure 18. Examples of four binary phantoms manually generated for the analysis of the newly developed shape descriptors. Regular and round (A), elliptical (B), high spiculated (C), and irregular (D) shapes are illustrated. These phantoms show different degrees of mass irregularity and simulate the common morphological characteristics of benign (A, B) and malignant (C, D) breast lesions.

3.2 Image Cropping and Data Augmentation

All the processes of image processing, function development, and data analysis were carried out in MATLAB (version R2018a, The MathWorks Inc., Natick, MA).
Each tumor was initially cropped into a 128 x 128 pixels patch (corresponding to 3.5 x 3.5 cm^2) to solely focus attention on the suspected lesion and reduce the computational cost necessary to process the images. This patch was built in the coronal plane encompassing the mass centroid, calculated from the segmentation volume. Its dimensions were the result of a size analysis of the masses: actually, all the lesions had a size to be completely enclosed within the patch area.

Considering the ML classification tasks and their dependency on the dataset size, the dataset available in this study was limited and would not be suitable in a radiomic-based approach, which adopts a large number of features. Therefore, for each mass, a higher number of 128 x 128 pixels patches were collected. Since DBCT is an isotropic tomographic imaging modality [39], 9 different views corresponding to the 9 planes of symmetry of an imaginary cube, built around the centroid of the mass, were extracted. These views were the coronal, the sagittal, the axial, and the 6 diagonals that contain two opposite edges and four vertices (Figure 19).

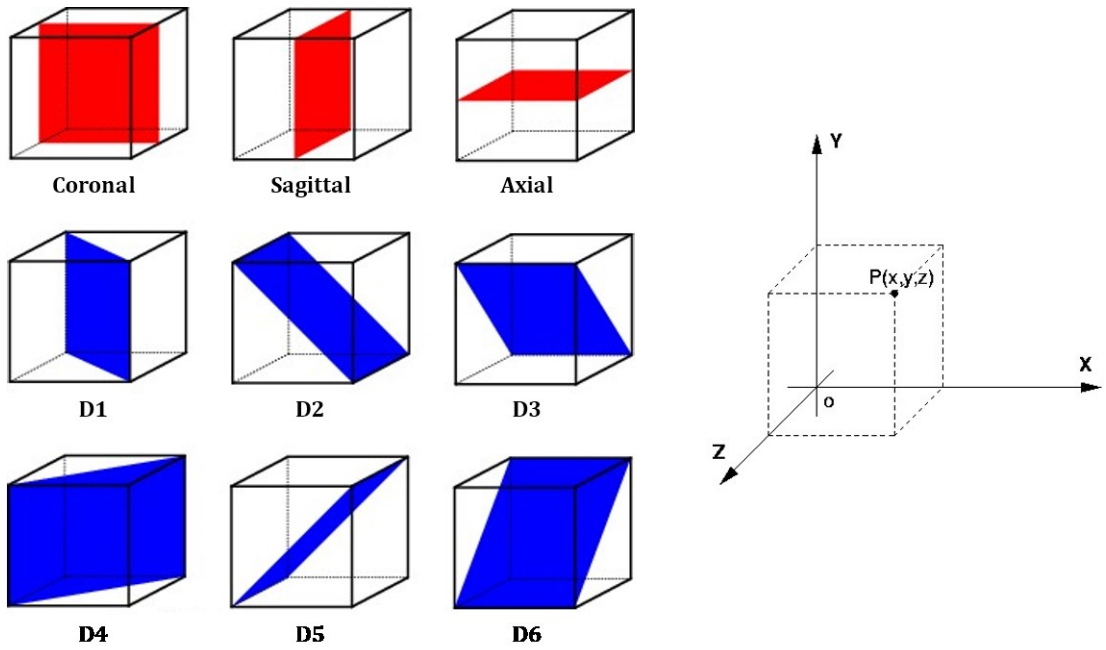


Figure 19. Planes of symmetry of a cube. On the left side, the three planes parallel to a pair of opposite faces (i.e. Coronal, Sagittal, Axial) are illustrated in red, while the six diagonal planes containing a pair of opposite edges and four vertices are in blue. On the right side, the Cartesian coordinate system used to define the rotations to be performed for the different views is shown. Retrieved from [74].

This augmentation process makes it possible to have more patches available starting from the same limited dataset and to capture the characteristics of the masses from different points of view but with the same spatial resolution.
It allowed obtaining more robust performance, preventing the overfitting that could have resulted if only coronal views had been adopted (see Section 3.4). Indeed, the multi-view analysis considered all the directions in which the suspect mass develops, without neglecting any traits that it might have on different sides.

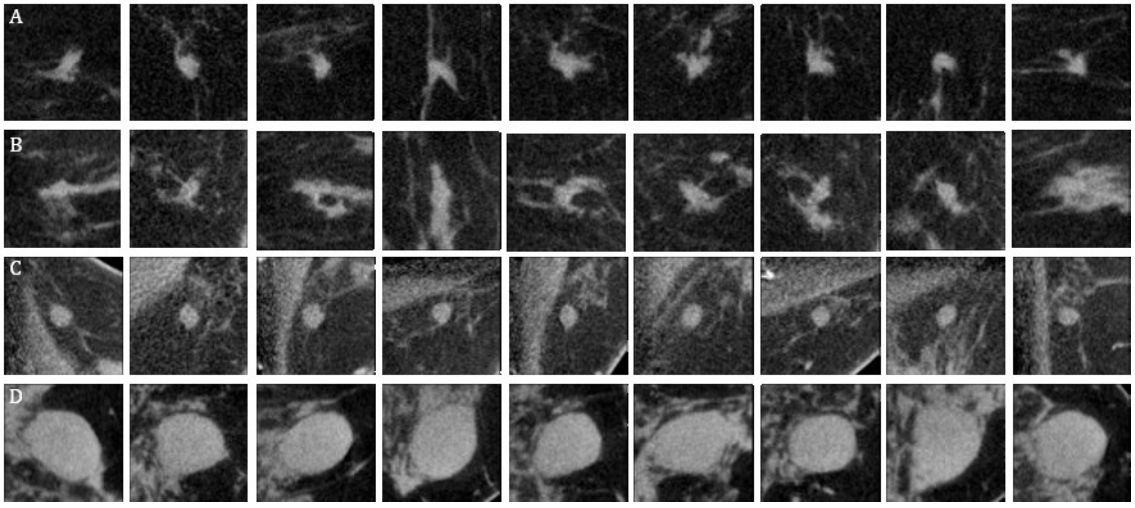


Figure 20. Examples of image patches generated with the data augmentation process. Each row illustrates the 9 different views corresponding to the same breast mass. In particular, the first two (A, B) refer to malignant masses, while the last two (C, D) to benign ones. Figure courtesy of Marco Caballo and Ioannis Sechopoulos.

From a technical point of view, the original volume that contained all the coronal slices was used, and the relative rotations were performed so that it was possible to collect the additional 8 patches from the different planes of symmetry (Table 1). Each patch was correctly centered and extracted from both the original and the annotated volume. A slight image processing was done to the segmentation patches to improve the quality of the binary mask (i.e. slight opening, hole filling, and removal of small external objects).

Table 1. Planes of symmetry and relative rotations generated to obtain all the patches collected from each mass of the dataset. The Coronal plane is the reference plane to obtain its orthogonal planes (i.e. Sagittal and Axial planes) and 4 diagonal planes (i.e. D1, D3, D4, D6), while the Sagittal plane is the reference for the two remaining diagonal planes (i.e. D2, D5). Refer to Figure 19 for the planes' name and the coordinate system.

Plane of Symmetry	**Rotation**
Coronal (X-Y plane)	-
Sagittal (Y-Z plane)	90° around the Y-axis of the Coronal plane
Axial (X-Z plane)	90° around the X-axis of the Coronal plane
D1	45° around the Y-axis of the Coronal plane
D2	45° around the X-axis of the Sagittal plane
D3	45° around the X-axis of the Coronal plane
D4	135° around the Y-axis of the Coronal plane
D5	135° around the X-axis of the Sagittal plane
D6	135° around the X-axis of the Coronal plane

The dataset of 74 masses was divided into training, validation, and test set, necessary to train, tune, and assess the ML model used for the diagnostic classification between benign and malignant tumors. The lesion partition was made to stratify these masses. Indeed, all types of tumors are evenly distributed over the three different sets, which were homogeneous in terms of the characteristics expressed and analyzed in the feature extraction process.

Two other datasets composed of a higher number of patches than that of only the extracted coronal ones were built to study the overfitting problem, which will be described in Section 3.4 in more detail: for now, *overfitting* is defined as the excessive fitting of the ML model to the training set used in the learning phase. In particular, the first of the two new datasets examined the three fundamental planes of the human body (i.e. Coronal, Sagittal, Axial) and included 222 patches, while the second considered all the nine previously described views and contained 666 samples. Following the initial subdivision in training, validation, test set made for the first original dataset, the patches of the two new datasets were distributed in the three sets so that the ones belonging to the same mass were part of the same set: for the sake of clarity, all the patches of a specific mass fell in either the training or the validation or the test set.

Therefore, although the patches refer to different views of the same mass and have no geometric reference that links them to each other (but only the lesion from which they are extracted), bias evaluation problems due to the presence of patches of the same mass distributed in the three sets were avoided.
The partition of the masses, and the respective patches, in the three different size datasets, as well as their further assignment in the training, validation, and test sets are reported in Table 2.

Table 2. Overview of the image datasets used in this study. Each of them presents the number of benign and malignant patches assigned to the three sets needed for the training, validation, testing stages of the ML classification model.

Dataset		Benign	Malignant	Total
Coronal Plane	Training	36	11	47
	Validation	4	2	6
	Test	14	7	21
Anatomical Planes (Coronal, Sagittal, Axial)	Training	108	33	141
	Validation	12	6	18
	Test	42	21	63
9 Planes	Training	324	99	423
	Validation	36	18	54
	Test	126	63	189

3.3 Shape and Margin Feature Extraction

All the radiomic features developed in this thesis are presented in this Section.
As announced in Section 1.4, the CADx system to be designed adopted only descriptors related to the shape and the margin of the breast lesions. Each of the two categories is described in a dedicated subsection. Novel and more complex descriptors will be provided with an exhaustive description and a mathematical definition within the feature group to which they belong.

3.3.1 Shape and Contour Descriptors

Shape and contour descriptors are the first to be taken into account and include four different feature sets, which examine the basic morphological properties, the centroid distance function, the region boundary, the spiculae and lobes of the mass, respectively. Their purpose is to extract information starting from the binary masks that report the manual segmentation of the lesions. A total of 28 shape descriptors were developed and are described below.

Basic Morphological Features

The basic morphological features are all descriptors easy to computational implement that can be found in most of the previously proposed radiomic pipelines. Each of the following paragraphs provides a brief description of the features in question. Although many of them are known as properties of the continuous domain, their definition evaluates the discrete domain because images are technically matrixes of finite elements (i.e. pixels).

Area is defined as the number of pixels present within the binary segmentation mask and measures the surface area of the lesion region. While considering the same overall dimensions, the area of regular and spiculated lesions is higher than that of irregular shapes, characterized by lobes and concavities.

Perimeter is referred to as the number of pixels that belong to the region boundary, which is the set of all those pixels lying between the segmentation mask (one-indexed pixel) and the unmarked part (zero-indexed pixel). This descriptor is returned as the sum of all the distances between the various pixels of the contour itself. Under the same dimensions, the perimeter of spiculated lesions is longer than that of regular and macro-lobulated ones.

Convex Area is the number of pixels belonging to the smallest convex polygon that allows enclosing all the tumor region area. Indeed, all the pixels of the lesion are contained within the *hull*, which is exactly the convex region boundary.

Still referring to the convex polygon that surrounds the tumor region, **Convex Perimeter** is the sum of all the distances of the pixels that constitute the hull itself.

Major and Minor Axes are the descriptors related to the pixel-based measurement of the major axis and the minor axis of the tumor segmentation mask. In particular, they calculate the length (in pixels) of the two axes starting from the ellipse which has the same normalized variance of the lesion region. Regular shapes are characterized by similar (and even identical) values, while the irregular ones can result in values that differ by some pixels.

Eccentricity is calculated starting from the previously mentioned ellipse. It is defined as the ratio of the distance between the two ellipse foci and the major axis length. Its value ranges between 0 and 1, where the degenerate cases correspond to the circle and the line segment, respectively. Therefore, this descriptor makes it possible to measure the degree of elongation of the region and how much it differs from a perfect circle: indeed, masses with irregular shapes (lobes, protrusions, concavities) show higher values than the regular ones.

Equivalent Diameter [75] and **Roundness** [57] measure the circularity of the lesion region and allow to discriminate regular masses from all the others. The former returns the diameter of a circle having the same surface area as the lesion under analysis and generally gives high value for regular and macro-lobulated masses. The latter assumes the maximum value of 1 if the shape is a perfect circle, and is typically high for rounded and regular tumors, medium for macro-lobulated ones, and low for spiculated ones. They are defined as:

$$\text{EqDiam} = \sqrt{4 \cdot \frac{\text{Area}}{\pi}} \tag{1}$$

$$\text{Roundness} = \frac{4 \cdot \pi \cdot \text{Area}}{\text{Perimeter}^2} \tag{2}$$

Solidity [76] is defined as:

$$\frac{\text{Area}}{\text{Convex Area}} \tag{3}$$

It measures the amount of convexity in the lesion region. It could be defined as an evaluation of *density*, which most characterizes benign lesions, given their typical regular shape. Indeed, the convex area is comparable to that of the tumor region in case of high circularity, and the value is usually high for regular and rounded masses. Irregular and spiculated lesions have a lower value due to the difference between their real area and the convex one.

Finally, **Convexity** [77] is defined as:

$$\frac{\text{Perimeter}}{\text{Convex Perimeter}} \tag{4}$$

It is the other side of *Solidity* because this descriptor has the same expression but with perimeters instead of areas. It numerically provides higher values for malignant tumors, characterized by spiculations and irregularities.

Centroid Distance Function (CDF) Features

The Centroid Distance Function (CDF) is the function that collects all the distances of the contour pixels of the lesion from the centroid of the region itself. There are different distance metrics, but, in this study, the Euclidean distance was adopted.
If the coordinates of each pixel are (x_i, y_i) and of the centroid are (x_c, y_c), the CDF referred to the i-th contour pixel is defined as:

$$CDF_i = \sqrt{(x_i - x_C)^2 + (y_i - y_C)^2} \tag{5}$$

CDF was calculated starting from the binary segmentation mask available for each tumor: it required both the centroid calculation and the identification of all the boundary points. Then, it was normalized by the maximal value of it, which also constitutes one of the parameters included in the feature set. This function plays a critical role in understanding the irregularity of the tumor shape because its profile clearly outlines the frequency and magnitude of spiculae, lobes, irregularities.
Some examples of real tumors belonging to the dataset and the respective extracted CDF are shown in Figure 21.
Taking into consideration these examples, breast masses with regular shape and contour were characterized by CDFs with a low-frequency and high-magnitude curve. Indeed, beyond the noise due to the discretization of medical images, the CDF waves were smooth. Moreover, since most of the contour pixels had the same distance from the centroid, the magnitude was high in almost all the pixels, presenting only two local minima in correspondence to the two points closest to the centroid (shortest radii). Instead, more spiculate masses had, on average, a higher frequency and a smaller module because of their indented contour and their less regularity in shape. Finally, more lobulated and irregular lesions exhibited waves with a high frequency and a considerable reduction in magnitude due to a higher number of ups and downs of the curve itself: this is because there were points at a higher or lesser distance from the centroid of the region. Therefore, from these CDFs, it was possible to evaluate the main morphological characteristics and use them as effective biomarkers for tumor diagnosis.
The proposed radiomics pipeline included 14 CDF-based descriptors, which are presented below.

Mean (m_1) and **Standard Deviation (M_2, STD)** constitute a first analysis of the CDF and offer an average interpretation starting from all the contour pixels on which CDF was evaluated. In particular, mean took higher and STD lower values for the more regular shapes, while mean decreased and STD increased when the mass became more irregular. Considering the spiculated lesions, mean and STD fell within the range of the two categories previously described because their mean and STD slightly decreased and increased, respectively. They are defined as:

$$m_1 = \frac{1}{N}\sum_{i=1}^{N}[\,CDF(i)\,] \tag{6}$$

$$M_2 = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[\,CDF(i) - m_1\,]^2} \tag{7}$$

Max Radius and **Min Radius** [75] are defined as:

$$MaxR = \max(CDF) \tag{8}$$

$$MinR = \min(CDF) \tag{9}$$

They identify the maximum and minimum Euclidean distance of the points that constitute the edge of the region from the center of mass of the region.

Elongatedness [75] is defined as:

$$\frac{Area}{(2 \cdot MaxR)^2} \tag{10}$$

This metric discriminates well the benign masses that are usually regular (high values) from the malignant ones that are spiculated and irregular (low values).

Dispersion [75] is defined as:

$$\frac{MaxR}{Area} \tag{11}$$

It measures the irregularity of the region based on the ratio between the maximum radius and the area. Indeed, lesions with a regular shape were characterized by low dispersion values, while its values increased if the number of concavities and irregularities grew.

Shape Index [75] is defined as:

$$\frac{\text{Perimeter}}{2 \cdot \text{MaxR}} \tag{12}$$

It provides information on the shape starting from the boundary of the region, since it derives from the perimeter and maximum radius. Spiculated shapes, like those of some malignant lesions, exhibited high values given their extended perimeter, while more regular shapes (e.g. round or with large lobes) had low values.

Area Ratio [78] is defined as:

$$\text{AR} = \frac{1}{m_1 \cdot N} \sum_{i=1}^{N} [\, \text{CDF(i)} - m_1] \tag{13}$$

Area Ratio (AR) quantifies the percentage of mass that lies outside the circular region having the CDF mean as its radius. Therefore, it describes the morphological characteristics of the tumor. It assumed small values for the masses that are almost circular: their mean value was close to the distance of the various boundary points from the centroid. AR value increased as the irregularity incremented, obtaining higher values for irregular and macro-lobulated lesions. On the other hand, spiculated masses were defined by an AR value included between those of regular shapes and those with macro-lobes.

Entropy [78] is the descriptor that expresses the disorder present within a given distribution. This study assesses the degree of randomness of the CDF, starting from the histogram built with the probabilities that each CDF element was within a specific bin defined between a length R and R+ΔR. Indeed, it is possible to evaluate the CDF diversity, determining how much the lesion shape under analysis is regular or not. Since it is an irregularity criterion, rounded masses had low values, while lobulated and spiculated ones were identified by high values. If the probability of falling into a given i-th bin is p_i, this radiomic descriptor is defined as:

$$\text{En} = \sum_{i=1}^{100} |p_i \cdot \log(p_i)| \tag{14}$$

F_1 Moment [79], **F_3 Moment** [79], **F_{31} Moment** [79], and **F_{3k} Moment** [80] are four low-order moments that analyze the irregularities present on the boundary. In particular, they are obtained as combinations of the mean and the STD parameters based on the CDF. Unlike high-order moments that are quite sensitive to noise, they proved to be able to discriminate the shape types present both in the original dataset of breast masses and the phantom dataset. These moments can provide an in-depth frequency analysis of the CDF, describing the relationships that exist between the boundary pixel distances and its elementary parameters, independently of scaling and rotation. They assumed lower values for regular shapes, while they became higher as the contour irregularity increases. Macro-lobulated lesions showed the highest values of these descriptors. They are defined as:

$$F_1' = \frac{[\,M_2\,]^{\frac{1}{2}}}{m_1} \tag{15}$$

$$F_3' = \frac{[\,\frac{1}{N}\sum_{i=1}^{N}[\,CDF(i) - m_1\,]^4\,]^{\frac{1}{4}}}{m_1} \tag{16}$$

$$F_{31} = F_3' - F_1' \tag{17}$$

$$F_{3k} = \frac{\frac{1}{N}\sum_{i=1}^{N}[\,CDF(i) - m_1\,]^4}{M_2} \tag{18}$$

The last descriptor associated with CDF is a parameter calculated starting from the CDF Discrete Fourier Transform (DFT): it takes the name of **Energy of Fourier Coefficients applied to CDF (FD_{energy})** because it evaluates the energy content of the power spectrum. The first step requires the DFT computation of the CDF vector, which is already normalized by its maximum value.

$$a_n = \frac{1}{N}\left|\sum_{i=0}^{N-1} CDF_i \cdot \exp\left(\frac{-j2\pi ni}{N}\right)\right| \qquad n = 0, 1, \dots, N-1 \tag{19}$$

Subsequently, min-max normalization of the Fourier coefficients is performed to make them invariant to scaling. Finally, the first Fourier descriptor corresponding to the zero frequency is set to 0 to make the Fourier series even independent of the initial boundary pixel position where CDF is extracted. At the end of the process, the series of normalized Fourier coefficients takes the following form:

$$C_n = \begin{cases} 0; & n = 0 \\ 2 \cdot \dfrac{a_n - \min(a)}{\max(a) - \min(a)}; & n = 1, 2, \dots, N/2 \end{cases} \tag{20}$$

FD_{energy} is calculated starting from these coefficients by the following expression, which takes into account the two-sided nature of the power spectrum:

$$FD_{energy} = \sum_{n=1}^{N/2} C_n^2 \tag{21}$$

The evaluation of the performances was accomplished both on the real masses and on the manually generated phantoms. Considering Figure 21 and Figure 22, regular and rounded masses hold power spectra concentrated towards the low frequencies since the corresponding CDF was smooth and dampened: FD_{energy} highlighted the low global energy content, assuming a low value. Instead, spiculated masses had noisier and higher frequency CDF, which was reflected in an extended power spectrum across the entire frequency axis. Therefore, their global energy was larger because it was contained at more frequencies, and FD_{energy} gained a high value.
Finally, macro-lobulated and irregular masses had an intermediate behavior: the global energy content was higher than the regular but lower than the spiculated ones, which resulted in values that were included in the range defined by the other two types of shapes. Moreover, what emerged from the phantom study carried out in Figure 22 was that this radiomic descriptor is independent of the rotation and size of the masses, an important aspect considering the numerous types of tumors that can be identified.

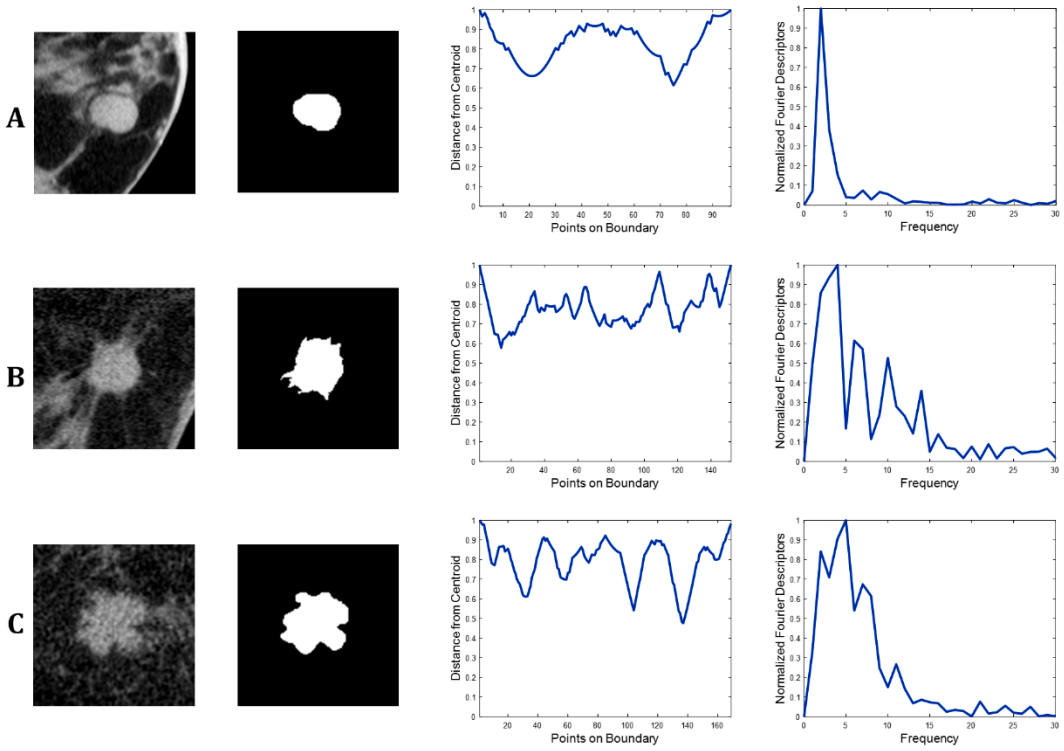
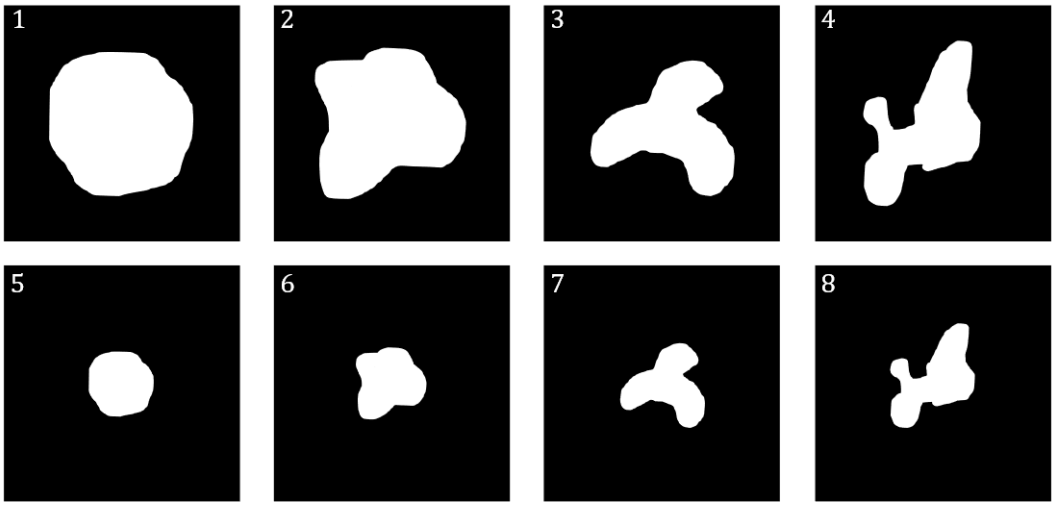


Figure 21. Examples of three different breast lesions belonging to the dataset. Each row refers to different masses: benign with a regular shape (A), malignant with spiculated contour (B), malignant with a macro-lobulated and irregular shape (C). For each of them, from the left to the right, the

original image patch, the binary segmentation mask, the Centroid Distance Function (CDF), and the CDF power spectrum are shown. Regular profile exhibits smooth and low-frequency CDF: its power spectrum shows a peak on the left side of the frequency domain. As the irregularity increases, CDFs reveal more rough, noisy, and high-frequency profiles, which are confirmed by the associated power spectra: in particular, the spiculated mass has a spectrum distributed along almost the entire frequency axis, while the macro-lobulated one has a spectrum with intermediate characteristics between the regular and the spiculated ones.



Mass	#1	#2	#3	#4	#5	#6	#7	#8
FD_{energy}	3.06 e-3	47.7 e-3	184 e-3	251 e-3	3.22 e-3	41.7 e-3	180 e-3	243 e-3

Figure 22. Examples of manually simulated phantoms for the validation of the Energy of Fourier Coefficients (FD_{energy}) radiomic descriptor, which is related to the global energy content of the Fourier power spectrum obtained from the centroid distance function (CDF). These eight phantoms are arranged in pairs to show its scaling invariance. From left to right, there are two regular, two macro-lobulated and two irregular masses. FD_{energy} is useful in discriminating different shapes since it assumes different values for the shape types being analyzed: its value increases as the irregularity increases and is size-independent.

Region Boundary Descriptor

This radiomic descriptor directly analyzes the pixels that constitute the lesion boundary to assess the frequency energy content associated with it [79]. Unlike FD_{energy} that is calculated on the CDF, this parameter derives from the application of the DFT to the boundary pixels. In particular, the first step to obtaining the Region Boundary Descriptor (RBD) involved the conversion of the Cartesian coordinates of the boundary pixels into their complex form. By identifying with (x_i, y_i) the Cartesian coordinates of a boundary pixel and with z_i its complex sequence, the expression of each i-th boundary pixel is written as:

$$z_i = x_i + j \cdot y_i \tag{22}$$

The second step is the DFT application to these N complex elements, where N is the total number of boundary pixels. The Fourier coefficients are obtained as:

$$A_n = \frac{1}{N}\left|\sum_{i=0}^{N-1} z_i \cdot \exp\left(\frac{-j2\pi ni}{N}\right)\right| \qquad n = 0, 1, \dots, N-1 \tag{23}$$

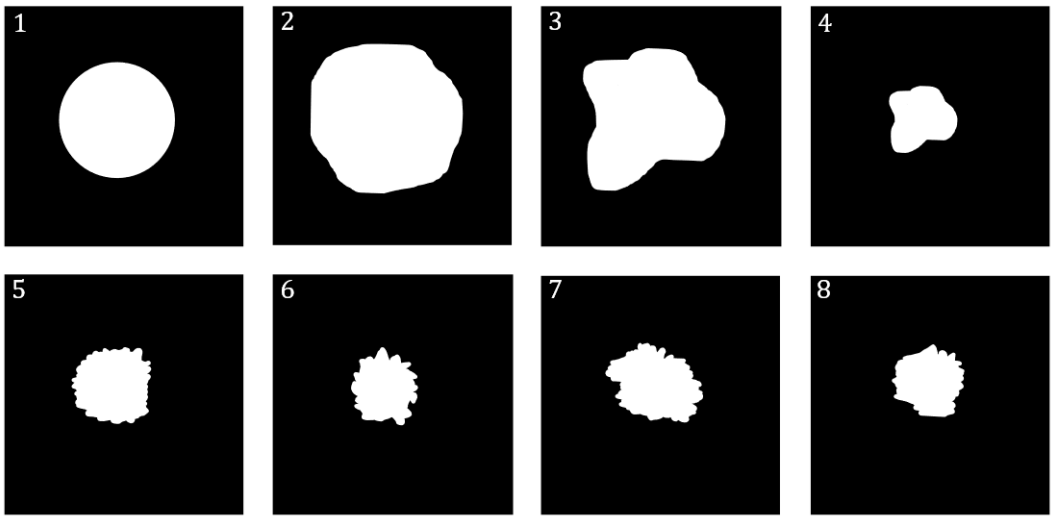
As done for FD_{energy}, the first Fourier descriptor corresponding to the zero-frequency is set to 0 to make the series independent of changes in the initial position, while the other coefficients are normalized with respect to the magnitude of the first non-zero frequency coefficient to make the Fourier series size-independent. Therefore, the Normalized Fourier Descriptors (NFD) have the following magnitude formulation:

$$NFD_n = \begin{cases} 0; & n = 0 \\ |A_n/A_1|; & n = 1, 2, \dots, N/2 \\ |A_{n+N}/A_1|; & n = -1, -2, \dots, -N/2+1 \end{cases} \tag{24}$$

RBD metric is defined as:

$$RBD = \frac{\sum_{n=-N/2+1}^{N/2} NFD_n/|n|}{\sum_{n=-N/2+1}^{N/2} NFD_n} \tag{25}$$

The choice to normalize all NFDs by the corresponding n-th frequency was made to increase the importance of the low-frequency components and contribute to the diagnostic discrimination of benign and malignant masses. Indeed, if the lesion has a spiculated and rough contour, RBD will be smaller because most of the energy is distributed across the high frequencies and is damped by these frequencies that have a large magnitude. Instead, tumors with regular and rounded boundaries have the global energy boxed in the low frequencies, which consequently will not tone down the contribution of the NFDs because of their small magnitude.



Mass	#1	#2	#3	#4	#5	#6	#7	#8
RBD	0.862	0.820	0.701	0.723	0.628	0.652	0.681	0.673

Figure 23. Examples of manually simulated phantoms for the validation of the Region Boundary Descriptor (RBD), which is related to the frequency energy associated with the Fourier spectrum obtained from the coordinates of the boundary pixels. These eight phantoms reconstruct the main contour characteristics of typical breast tumors: regular and rounded (1, 2), macro-lobulated (3, 4), spiculated and rough (5, 6, 7, 8). What emerges from the results of the study, RBD decreases its value when the contour becomes more irregular. The motivation is linked to the mathematical formulation of the radiomic descriptor, which provides a division of each Fourier coefficient by the magnitude of the respective frequency: it means that masses with a global energy content localized at low frequencies (i.e. benign tumors) have higher RBD values.

Spiculae and Lobes Map (SLM) Features

Spiculae and Lobes Map (SLM) descriptors constitute a set of features developed from scratch in this thesis, taking inspiration from the study of Kpalma et al. [81] applied in the context of pattern analysis and recognition. They perform an analysis of the irregularities of the mass conformation in terms of spiculae, lobes, and concavities. The necessary elements for the calculation of these radiomic features are the original contour of the lesion and its convex enveloping curve, which outlines the smallest convex area containing all the pixels that define the binary segmentation mask. These parameters are obtained from the intersection between the original contour and the convex enveloping curve, which is iteratively eroded using a circular morphological structuring element with a radius that increases as the iterations progress. The process stops when there are no further intersections between the two curves. Some examples related to the real breast lesions belonging to the original dataset are shown in Figure 24.

The descriptors initially extracted are the maximum number of intersection points between the two curves on a single iteration ($SLM_{intersections}$), and the maximum number of iterations executed before the erosion process interruption ($SLM_{iterations}$). These two features contain information that has not been analyzed so far to evaluate the degree of spiculation and the number of inflections present within the lesion. The two descriptors proposed in the radiomics pipeline of this study are the result of the combination of the previously mentioned parameters, actually extracted from the region of interest. In particular, the final metrics are precisely the product ($SLM_{product}$) and the ratio (SLM_{ratio}) of the two quantities, and are presented below:

$$SLM_{product} = SLM_{intersections} \cdot SLM_{iterations} \tag{26}$$

$$SLM_{ratio} = \frac{SLM_{intersections}}{SLM_{iterations}} \tag{27}$$

From the analysis carried out both on the original masses and manually simulated phantoms, regular shapes are characterized by a low number of intersections and a low number of iterations, which resulted in a small $SLM_{product}$. In contrast, spiculated masses have a high number of irregularities and asperities, which came out in a high number of intersections but a low number of iterations since they are only on the marginal side of the lesion and are not excessively deep. Therefore, $SLM_{product}$ and SLM_{ratio} were both high since the product is linearly dependent on the number of intersections, and the ratio had a small number of iterations in its denominator. Finally, irregular and lobulated masses have concavities and lobes that are fewer in number than the previous group, but they are deeper. Thus, $SLM_{product}$ maintained a high value because of the number of iterations, while SLM_{ratio} dropped due to its large denominator.

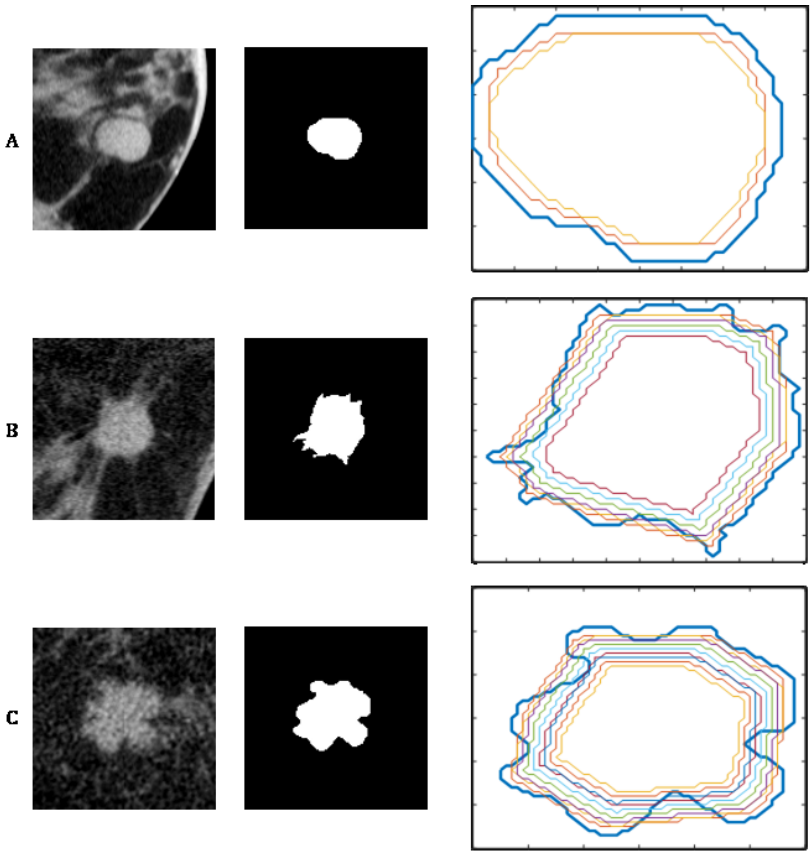


Figure 24. Examples of three real breast lesions from which the SLM features were extracted. Each row refers to different masses: benign with a regular shape (A), malignant with spiculated contour (B), malignant with a macro-lobulated and irregular shape (C). From the left to the right, the original image patch, the binary segmentation mask, and the erosion process of the convex enveloping curve performed until there are no further intersections between this curve and the original mass contour, are shown. Regular masses exhibit few intersections and iterations, while spiculated and irregular ones have a higher number of them. The last two shape groups have different values of intersection and iterations due to the characteristics of their irregularities.

For SLM descriptors, a more detailed study was carried out to clarify the justification that led to not directly adopt the number of intersections and iterations but to determine combinations that are useful to discriminate benign and malignant masses based on their most typical and noticeable characteristics. Therefore, their diagnostic power was assessed through a phantom-based study. 900 digital mass phantoms representing the three groups of regular, irregular, and spiculated breast lesions were automatically generated. Each group consisted of 300 elements, equally divided into small, medium, and large size to test also the scaling invariance. These phantoms were generated from a perfect circular input on which a processed random white noise (zero mean, unit STD) was applied: by modulating the noise amplitude and varying the size of the moving-average filter kernel, the typical characteristics for each of the three groups were simulated: the ranges of the parameters related to the colored noise and the smoothing kernel are reported in Table 3. The different sizes of the phantom lesions were obtained using an initial radius of 25, 50, and 75 pixels for small, medium, and large masses, respectively.

Table 3. Noise amplitude and filter kernel size adopted for the automatic generation of the phantoms used in the design, development, validation, and testing phases of the SLM descriptors. The numerical values were defined to build phantoms that simulate the typical characteristics of the three groups of lesions under analysis (regular, irregular, spiculated). The values chosen for the noise amplitude and the filter kernel for the generation of the 900 phantoms (300 for each shape group) were randomly extracted within the declared ranges to obtain small differences for each simulated mass.

	Regular	**Irregular**	**Spiculated**
Noise amplitude [pixels]	3 – 5	15 – 20	9 – 12
Filter kernel size [pixels]	90 – 120	30 – 50	3 – 15

The results obtained with the two raw parameters ($SLM_{intersections}$ and $SLM_{iterations}$) and the two final ones ($SLM_{product}$ and SLM_{ratio}) are shown in Figure 25. The former contain the same information, distinguishing regular masses from the other two groups. The distributions of the three different groups show an overlap of the mean values that fall within the STD ranges of the others: indeed, irregular and spiculated distributions are substantially overlapped, emphasizing the difficult utility of the raw parameters adopted separately. Conversely, the latter hold a relevant position because $SLM_{product}$ provides clear discrimination between the regular masses and the remaining ones, while SLM_{ratio} well separates the group of spiculated masses from the rest. Although the STD ranges are partially overlapped, the positive aspect is that the mean values of the groups to be discriminated through each of the two derived parameters (regular with $SLM_{product}$, spiculated with SLM_{ratio}) are not contained within the STD ranges of the others. Despite the simplicity of their mathematical formulation, the two final SLM descriptors, proposed in a completely new way and not implemented in any radiomic pipeline, demonstrated during the development and validation phases a strong discriminative connotation for the identification of benign and malignant breast tumors starting from the quantification of spiculae, lobes, and concavities that characterize their shapes.

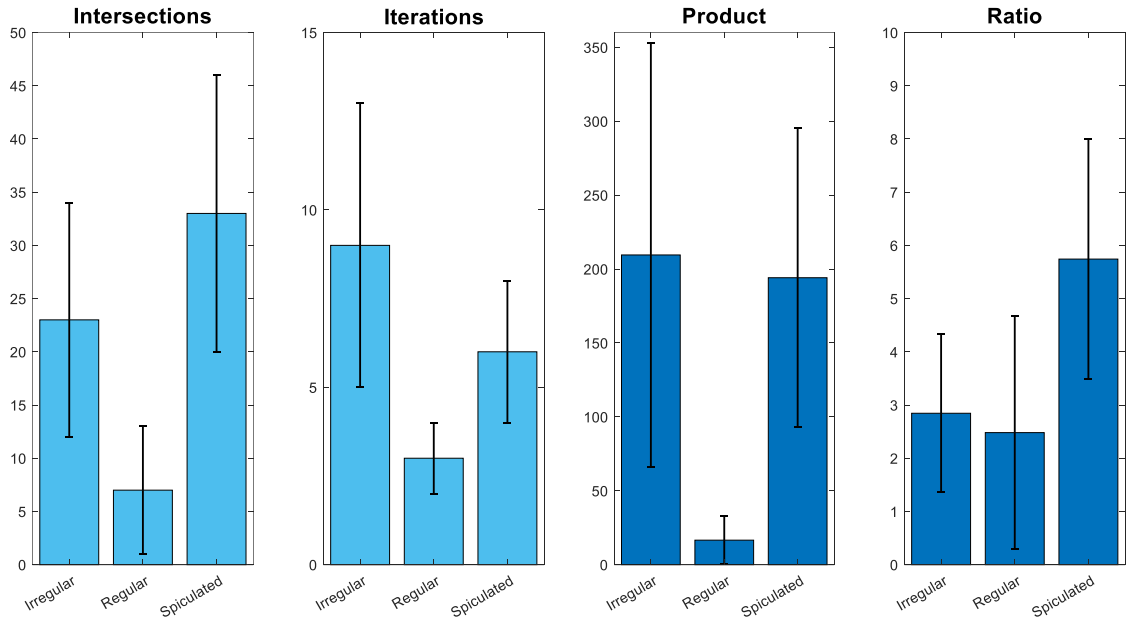


Figure 25. Results (mean, STD) obtained from the analysis of the SLM descriptors applied to the automatically generated phantom dataset. On the left (in light blue color), the distributions related to the two raw parameters ($SLM_{intersections}$ and $SLM_{iterations}$) are illustrated. They propose equivalent information regarding the differentiation of the three shape groups because only the regular one is separated from the others. Moreover, irregular and spiculated groups have a considerable overlapping of the STD ranges with the two mean values falling within the range of the other group. On the right (in dark blue color), the distributions related to the two derived parameters ($SLM_{product}$ and SLM_{ratio}) are shown. They are the final SLM descriptors because, if used together, are functional to the discrimination process since $SLM_{product}$ allows the regular masses to be distinguished from the others, while SLM_{ratio} facilitates the identification of the spiculated ones. Although partially overlapped, the mean values of the two groups to be discriminated are always out of the STD ranges of the other groups, giving further support to the usefulness of these radiomic descriptors.

3.3.2 Margin Descriptors

Margin descriptors explore the peritumoral compartments of the breast masses starting from the acquired gray-scale images and are divided into two feature groups, related to the analysis of the distribution of the radial gradient and the gray-level texture. The region being analyzed is the circular crown whose centerline is the contour of the binary segmentation mask. A total of 130 margin features were developed, and are explained below.

Margin Radial Gradient Distribution Features

These descriptors are designed to quantify the sharpness degree of the breast tumor margins. What can be observed starting from the gray-scale images is that malignant lesion margins are generally blurred and ill-defined due to the intricate micro-vessels network that typically surrounds these masses for blood supply and cancer proliferation. Instead, the benign lesions are commonly characterized by sharp and well-defined contours. Therefore, the idea was to define some margin indicators that analyze the edge gradient distribution to extract exploitable information for understanding the tumor status [82, 83].
As previously said, the lesion margin was identified as an annular region built on the contour of the segmentation mask. Its total thickness was set equal to 10 pixels, equally divided between inside and outside the ROI. Thereby, it was possible to evaluate the characteristics of the lesion both in its internal area and surroundings, developing a set of radiomic biomarkers that provide details about the intensity transition across the boundary. The gradient magnitude was calculated by the convolution of the image with the Sobel filter, particularly recommended for the border identification. Unlike the traditional extraction of texture features, the goal is to analyze the edge-gradient distribution within the margin region along the radial directions that connect each boundary pixel with the centroid of the lesion. Considering to have N contour points, there are N radial edge-gradient profiles from which 9 different descriptors are extracted: 8 first-order statistics features (mean, STD, maximum, minimum, energy, skewness, kurtosis, entropy) and full-width at half-maximum (FWHM). To associate a limited number of measurements with each lesion being analyzed, mean and STD are calculated for each distribution associated with one of the previously listed descriptors, defining 18 new radiomic metrics to be included in the pipeline of this study.

The mathematical expressions of the 9 extracted descriptors are presented below.

- **Mean**, defined as:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{28}$$

where x_i is one of the 10 pixels which constitutes the radial gradient profile contained within the annular region. It returns the profile average intensity.

- **STD**, defined as:

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{29}$$

It returns an estimation of the variability present within the gradient profile.

- **Maximum** and **Minimum** are the largest and smallest values of the radial edge-gradient profile, respectively. They represent the extremes of the intensity range within all the pixel values fall.

- **Energy**, defined as:

$$E_s = \sum_{i=1}^{n} |x_i|^2 \tag{30}$$

It is the energy associated with the gradient profile, calculated as the sum of the square modules of the intensity values.

- **Skewness**, defined as:

$$s = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\sigma^3} \tag{31}$$

It provides a measure of the distribution asymmetry of the pixel values around their mean.

- **Kurtosis**, defined as:

$$k = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\sigma^4} \tag{32}$$

It quantifies the distancing of the distribution from a normal distribution, giving an information related to the shape.

- **Entropy**, defined as:

$$S = -\sum_{i=1}^{n} P_i \cdot \log_{10} P_i \tag{33}$$

  where P_i is the intensity of the i-th pixel belonging to the radial gradient. This descriptor gives an average information regarding the profile being analyzed.

- **FWHM** is defined as the width in pixels that the intensity curve of the gradient profile has at the half of the maximum y-value assumed by the distribution. It is obtained as the difference between the x-coordinates of the points intersected at half the maximum amplitude.

These radiomic features were applied both to the real masses and some phantoms to evaluate their quantitative descriptive power in breast tumor discrimination. Considering what previously described, the benign margins are expected to be more homogeneous, while the malignant ones present a higher inhomogeneity and a greater degree of blurring. Figure 26 shows the 9 radial distributions applied to the same phantom to evaluate what changes in the presence of different mass sharpness degrees, while Figure 27 illustrates the application of two of these radiomic descriptors on real breast masses.

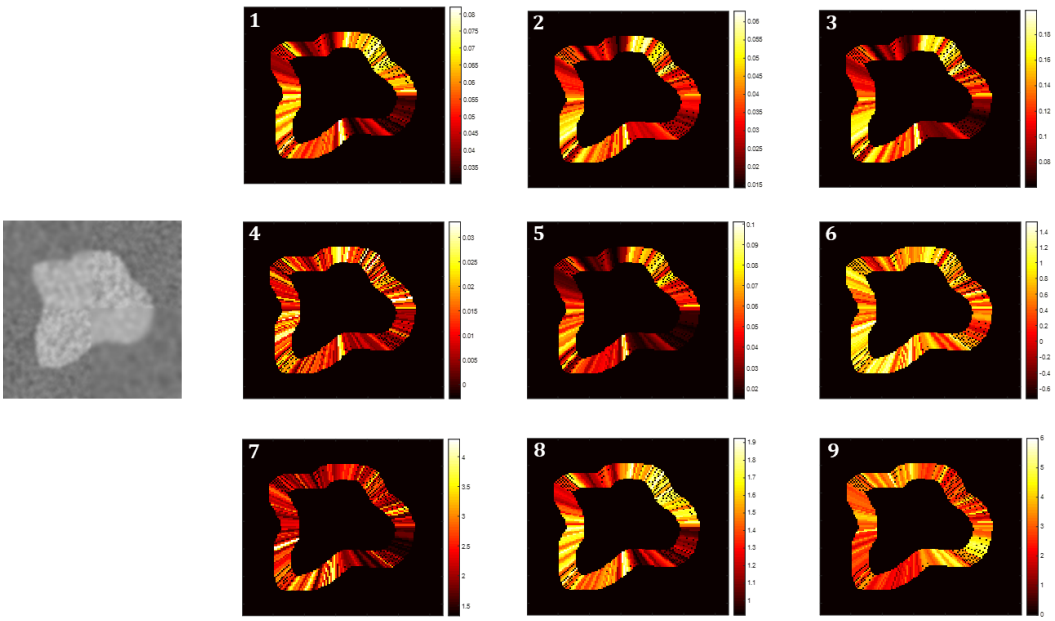


Figure 26. Phantom study of the nine margin radial gradient distribution features. On the left, the gray-scale phantom exhibits different sharpness degrees in its four quadrants: in particular, the 2D image-blurring filter was applied once in the upper-right and lower-left, twice in the upper-left, and three-times in the lower-right quadrants. On the right, the nine heating maps show the application of the nine descriptors to each radial edge-gradient profile of the mass phantom. In order: Mean, STD, Max, Min, Energy, Skewness, Kurtosis, Entropy, FWHM. All of them reveal different values depending on the blurring level, thus their mean and STD values add further useful information for the quantification of biomarkers related to breast tumors.

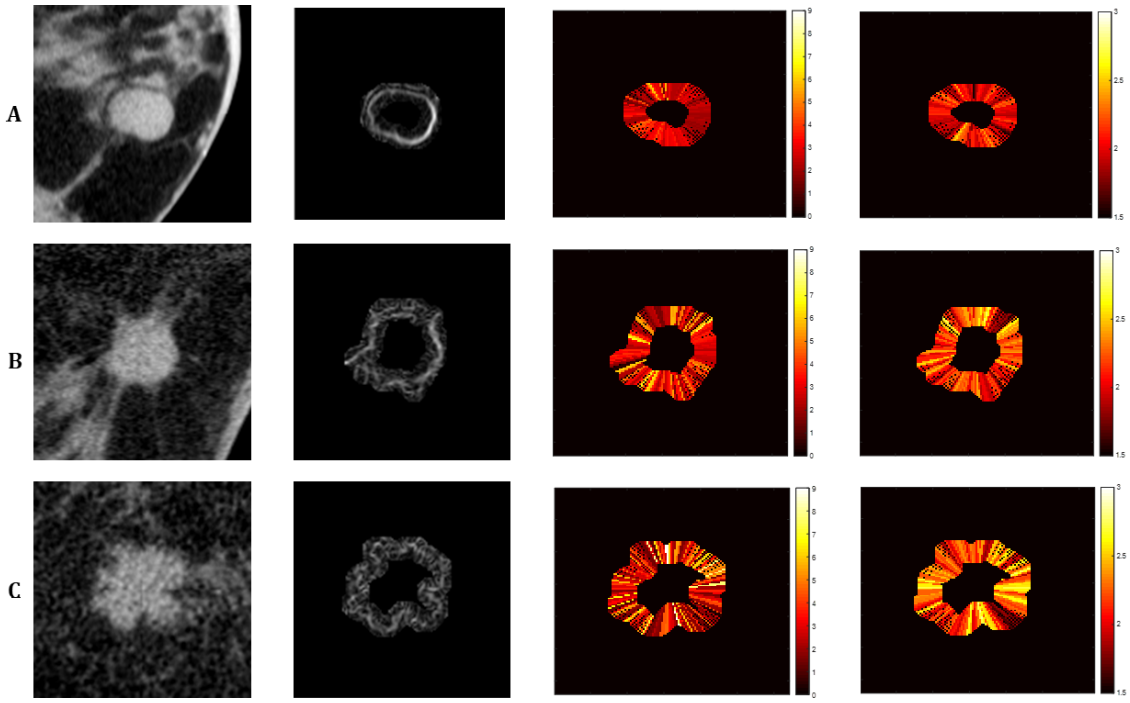


Figure 27. Examples of three real breast lesions from which the margin radial gradient distribution features were extracted. Each row refers to a different mass: benign with a regular shape (A), malignant with spiculated contour (B), malignant with a macro-lobulated and irregular shape (C). From the left to the right, the original image patch, the gradient of the lesion margin, and the heat maps of the FWHM and Entropy are shown. Benign masses usually show homogeneous heat maps in all the directions in which the radial edge-gradient is analyzed. Instead, malignant masses are characterized by inhomogeneous distributions, due to lower margin sharpness degrees and ill-defined margins.

By taking into account the FWHM descriptor, it results in a higher mean value for those masses that have more blurred margins and whose boundary line between the internal and external regions is less distinct, as in the case of malignant tumors: they yield radial gradient profiles that are a broad peak distribution. Therefore, if the lesion has a blurred margin in many orientations, the FWHM mean and STD will be higher because the gradient profiles will be different in each direction. For the sake of clarity, a small phantom-study on the same shape but with various sharpness characteristics is shown in Figure 28.

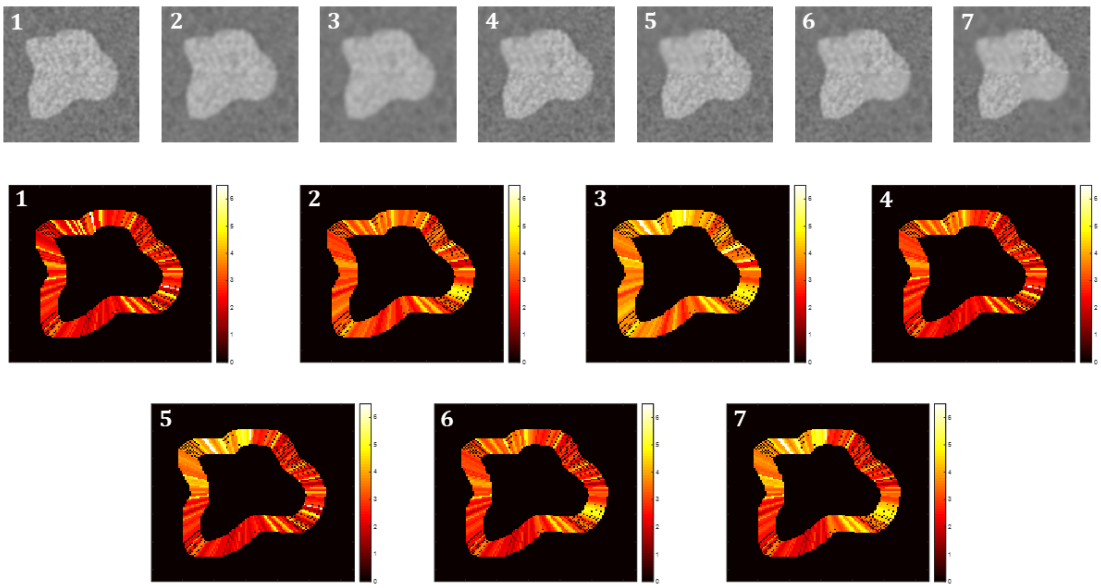


Figure 28. Examples of seven gray-scale mass phantoms with the same object but different blurring degrees from which the FWHM descriptor was extracted. The phantoms were obtained as follows: (1) was generated with random white noise and no filtering, (2) with a Gaussian blurring applied once over the whole (1), (3) with a second application of the blurring filter to (2), (4) with a blurring filter applied once to the upper-left corner of (1), (5) with a blurring filter applied twice in the upper-left corner of (1), (6) with a blurring application in the upper-left and lower-right quarters of (1), and (7) with a blurring application repeated twice in the upper-left and lower-right regions of (1). As can be seen, the heat maps show how sharpness variations lead to different values of the FWHM descriptor: the FWHM value becomes higher (yellow color) as the blurring content increases.

Margin Radial Sector Features

This set of radiomic descriptors analyzes the gray-level texture of the breast mass inside the annular region built as described in the previous subsection. Therefore, unlike the traditional texture features that are extracted from the entire mass, these margin metrics only describe the peritumoral regions to get useful information from a part of the lesion that contains many distinctive characteristics about the tumor nature [84, 85]. In this study, the circular crown was divided into ten sectors (one every 36°), and, from each of them, 14 features were extracted: 8 first-order statistics features (the same of the previous feature set) and 6 Haralick texture features (Contrast, Correlation, Energy, Homogeneity, Entropy, Symmetry [51]).
Eight first-order statistics features were calculated starting from the distribution of the 14 descriptors extracted from each of the ten sectors, to reduce the amount of information associated with each breast mass patch. A total of 112 margin radial sector features were implemented within the radiomic pipeline of this study. Some examples of the application of these descriptors (Haralick Contrast and First-order Energy) on real breast lesions are illustrated in Figure 29: these heat maps provide valid evidence of how each of the margin radial sector features can aid in the discrimination between benign and malignant mass images.

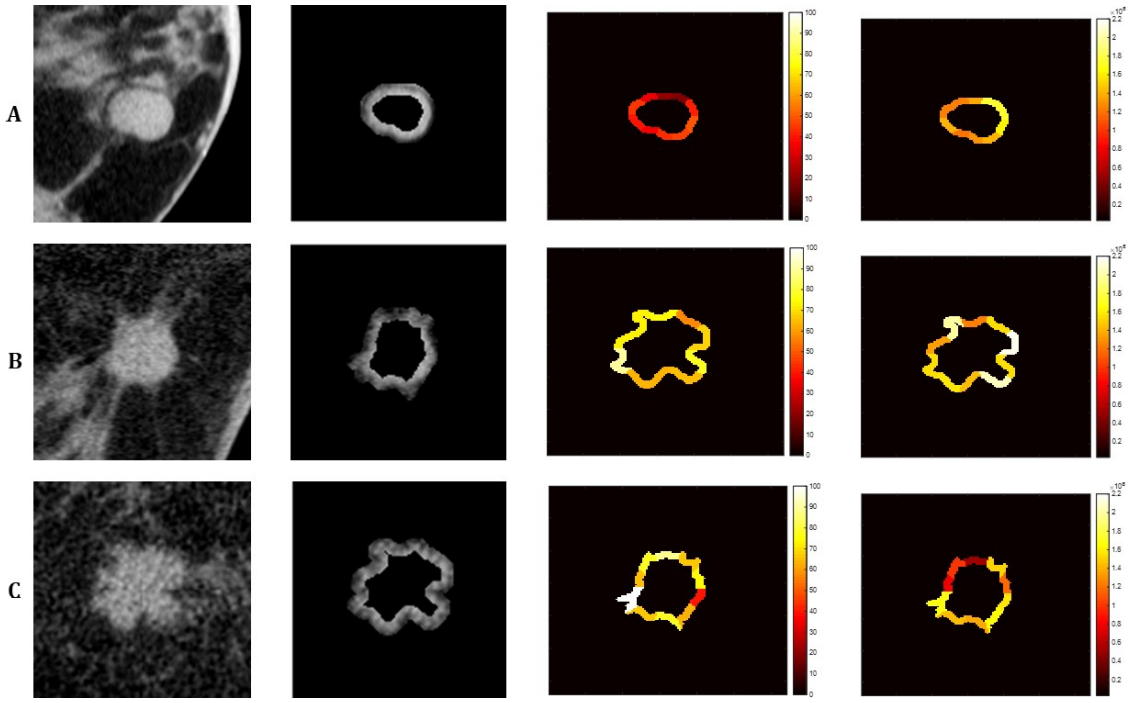


Figure 29. Examples of three real breast lesions from which the margin radial sector features were extracted. Each row refers to a different mass: benign with a regular shape (A), malignant with spiculated contour (B), malignant with a macro-lobulated and irregular shape (C). From the left to the right, the original image patch, the annular region of the mass margin, and the heat maps of the Contrast (Haralick) and Energy (First-order statistics) are shown. Benign masses usually show a certain homogeneity among the ten sectors, while malignant ones are characterized by sectors having significant differences in their characteristics (margin infiltration and ill-defined boundaries).

3.4 Machine Learning Analysis

All the Machine Learning (ML) analyses, methods, and models that were studied and implemented in this thesis are reported in this Section. In particular, the first subsection provides the ML definition and describes its basic concepts. The second subsection focuses on the overfitting problem and the feature selection techniques adopted in this study. Finally, the third subsection concerns the process of dataset balancing, which is critical for training ML classification models in the case of datasets with skewness in class proportion.

3.4.1 Machine Learning Introduction

ML is a branch of Artificial Intelligence (AI) and consists of algorithms, techniques, and methods that can learn a specific task (classification, regression, clustering, recommendation) from a set of data without being explicitly pre-programmed in all their components. Their performances are kept under analysis to understand if the achievement of their objectives is improved based on the knowledge learned, and the errors made during the training phase.
ML learning algorithms can be divided into two categories: *supervised* and *unsupervised*. In this thesis, a supervised ML algorithm was developed, so only a brief analysis concerning this approach is described.

Supervised ML refers to a type of learning that adopts a labeled training dataset, in which the training samples already have a tag that defines the group to which they belong. Therefore, the learning purpose is to extract patterns and rules that map the inputs (e.g. radiomic features) with the known output (e.g. benign vs. malignant), and, from it, produce a model that predicts the output of new samples. This approach is precisely the opposite of what happens in unsupervised learning, where the elements of the training set have no output labels, and, therefore, the training occurs by entirely devoting to the algorithm the task of finding possible patterns in the input data. Supervised ML involves a training phase, where the model can learn from the input characteristics of the training samples (both inputs and labels are provided), and a prediction (or validation/testing) phase, where the model predicts the label of new elements whose label is unspecified. In the validation and testing phases of the ML model, labeled input data are, therefore, only used to evaluate the model performance (i.e. such samples were not used during the training phase) and to understand if it has a correct behavior in determining the desired output in an out-of-sample evaluation. From a mathematical point of view, the training dataset in a supervised algorithm is defined as an $m \times n$ input matrix, called X, and a $m \times 1$ output vector, called Y, where m is the number of samples and n the number of input features extracted for each element.

If $x^{(i)}$ is the input vector that contains all the *n* features for a given training example, and $y^{(i)}$ indicates the output label, the training set (D) can be represented as:

$$D = \left\{\left(x^{(i)}, y^{(i)}\right)\right\}_{i=1}^{m} \tag{34}$$

The Y vector can contain discrete or continuous values, resulting in a classification or regression problem, respectively. The target variables of this study are binary (benign tumor 0, malignant tumor 1), and the goal is to build a breast mass classification model. In summary, the samples of the X matrix consist of all the patches containing the breast lesions (Sections 3.1 and 3.2), from which the 158 features were extracted (Section 3.3), and whose output can be benign (0) or malignant (1). To evaluate the performance of this ML binary classifier, before being included in a real CADx system for breast masses, this model was applied to the patches that compose the test set, to estimate its predictive power on a separated sample population.

3.4.2 Feature Selection

One of the most common problems encountered during the development and training of an ML model is *overfitting* [86]. Overfitting takes place when the ML model adheres too well to a particular set of data (training set) and is not able to generalize on new unseen samples, making incorrect predictions on a test set.
When overfitting occurs, prediction performance on the training data increase, while performance on never seen datasets get worse, and the generalization error becomes bigger. The disproportion between the number of samples belonging to the dataset and the number of input features is among the main reasons that lead to the overfitting problem. Indeed, if the attributes are more than the data samples, there is a high possibility of incurring in overfitting because having many features to deal with might need to force the building of a complex model that fits the training examples well, causing loss of generalization. One of the options to keep this issue under control is to reduce the number of features, which can also be very large in a radiomic-based approach: this process is called *Feature Selection* (FS). Likewise, overfitting can be prevented without performing any FS techniques if large datasets are available. Furthermore, one of the operations performed to avoid overfitting is to interrupt the learning process before the model fits the training samples excessively. There are many strategies for preventing overfitting, among which the *early stopping* can be mentioned. It involves the splitting of the original dataset in the training set and validation set, a set of data usually smaller than the former used to block the training when the error made on it starts to increase (Figure 30). Thus, this procedure constitutes another way to keep the overfitting problem under control, but it was not used in this thesis due to its difficulty in properly identifying the stopping iteration where the model overtraining begins.

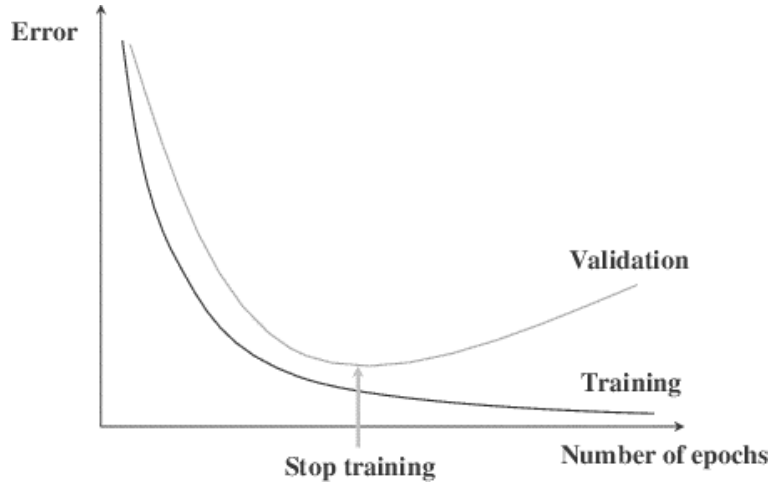


Figure 30. Early Stopping principle. The trend of the generalization error (y-axis) on the training and validation sets as the training process progresses (x-axis) is illustrated. There is a point (indicated by the blue arrow), where the error on the validation set begins to increase and where, therefore, the learning process should be interrupted to avoid overfitting the model on the training samples. Retrieved from [87].

It should not go unaddressed that FS has multiple advantages that concern both the overfitting prevention and the improvement of the ML model results [88]. Indeed, FS allows selecting the features that are considered most relevant and necessary to achieve the final objective, removing those that produce noise or that are uninformative, irrelevant, and redundant. The challenge is to identify the minimum subset of descriptors that can capture all the valuable information from the training set and find which intrinsic structures are present in the data used in the learning phase, to propose them when diagnostic conclusions need to be drawn. Therefore, FS covers a relevant role in the ML model building process because it ensures that only the most descriptive and significant attributes are part of the feature set to improve classification accuracy. All the FS methods implemented and tested in this study are described below.

Correlation

The Correlation-based Feature Selection (CFS) method generates the best feature subset based on the hypothesis that the selected descriptors are highly correlated with the classification goal and not correlated with each other [89]. Indeed, this approach is based on the analysis of the linear correlation between quantitative variable pairs, using the Pearson's linear correlation coefficient:

$$\rho_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \ \sum_{j=1}^{n}(y_i - \bar{y})^2}} \tag{35}$$

Where n is the variables' size, x_i and y_i are the i-th individual sample of the first and second variable, respectively, and $\bar{x}$ and $\bar{y}$ are the mean of the two variables. Its absolute value can range between 0 and 1, which indicate absence and perfect correlation, respectively. The choice of the threshold from which the features are considered highly correlated is arbitrary, and, in this study, it was set at both 0.8 and 0.9 (usually a $|\rho_{x,y}|$ above 0.7 is considered a strong correlation). When the linear correlation between the two descriptors is higher than the fixed threshold, one of the two is removed from the feature set. The higher the established threshold, the smaller is the number of discarded features because only the highly correlated descriptors are excluded.

To assess the significance level of the calculated correlation, the P-value is also considered for testing the no-correlation hypothesis: if the P-value is lower than the significance level of 0.05, then the corresponding correlation is statistically meaningful and can be evaluated to decide on whether or not to remove the descriptor.

Mann-Whitney U-Test

The Mann-Whitney U-Test is a nonparametric statistical test often used to verify whether two independent statistical samples come from the same population or not. In this study, the two groups were the set of benign and malignant patches. The goal is to understand which features are statistically significant to distinguish the two tumor classes, selecting only those descriptors that have a P-value lower than the level of significance set at 0.05. Therefore, FS is performed based on the discrimination power of each radiomic feature [90].

Mann-Whitney U-Test & Correlation

A third FS method adopted in this study is the combination of the two previously described methods. The results returned by the Mann-Whitney U-test are used to rank the features according to their importance, where the importance estimation is the P-value associated with their statistical test: the lower the P-value, the greater the importance of the feature. Therefore, after sorting the descriptors in ascending order (from the lowest P-value to the highest), the correlation-based method is applied. Under these conditions, when two variables are highly linear correlated with each other, the least important feature is discarded because it is ranked below in the feature hierarchy. The threshold is fixed only equal to 0.90.

Decision Tree Criterion (Predictor Importance)

Decision Tree (DT) is a flowchart-like tree structure that has become a widely-used supervised ML algorithm for classification tasks. The characteristic elements of a DT are the *root*, that is the topmost node from which the decision process begins, the *internal nodes* (also called *test nodes*), where a choice regarding a given variable is made, the *branches*, which constitute the outcome of the test nodes, and the *leaves*, which represent the terminal nodes and hold the class label of the element that is inside it. The dominant purpose for which DTs were born is to classify elements, after having trained the model with a labeled dataset. DT tends to over-train because the tree creation leads to adapt the characteristics of the graph to the elements of the training set. One of the approaches used for controlling the overfitting problem is to determine a priori the maximum number of leaves or the minimum number of training elements that can be present within each leaf: in this study, the minimum number of leaf node observations is 3, to avoid leaves made up of 1 or 2 instances. At each internal node, a feature is chosen as the best criterion to divide the data based on a test that allows obtaining the lowest impurity split. Therefore, it is possible to analyze the *predictor importance*, calculated as the sum of the Mean Squared Error (MSE) variations due to splits made by each variable divided by the number of internal nodes in which the variables are involved. This information is used to select the features that will constitute the final subset used to build the ML classification model. In particular, the FS method employed in this study selects all the features involved in the DT construction process, whose estimate of predictor importance is greater than 0.

Random Forest Criterion (Predictor Importance)

Random Forest (RF) is an ML supervised method used for classification tasks. It is part of the *ensemble* learning algorithms because it derives from the combination of several DTs. Indeed, it builds more DTs and puts them together to get a more stable and accurate prediction. In this study, the number of trained trees was equal to 200, and the minimum number of samples per leaf was set to 3 (as in the case of Decision Tree Criterion above). The goal of training together multiple DTs is to overcome the overfitting difficulties that arise when implementing an individual DT and build a model that maintains its generalization capabilities. The outputs of all the trained DTs are aggregated by majority voting in the classification task. To construct the M DT models, the training set is divided into M subset consisting of elements that are extracted from the original dataset by performing the *Bagging (Bootstrap aggregating)*, a uniform sampling with replacement method. Thus, it reduces the variance that characterizes the DT and adds a random contribution to the training process. Moreover, the RF method applies a selection of a random subset of features for each internal node: instead of considering all the variables for each split, some features are randomly chosen, and the best split predictor is selected only between

them. Actually, in this study, all the radiomic features were considered at each decision split, giving up the variability introduced by this second aspect and only adopting the Bagging method for the variance reduction. Another aspect on which this thesis focused is the algorithm chosen to select the best split variable at each node: Standard CART [91] and Curvature Test [92] were considered. The former is the most widely used technique, and selects the best splitting predictor that maximizes the gain of the split-criterion (*Gini's diversity index*). The latter selects the best splitting predictor that minimizes the P-value of the Chi-squared test between each predictor and its response.
The evaluation of the feature importance within the RF construction process requires the analysis of the impurity associated with each feature by using the out-of-bag samples, which are all those elements that are not included in a given training subset. As described above, the subset generation involves a random sampling with replacement, which means each sample could be extracted several times for the same subset. Thus, all those observations that are not part of a given subset are used to determine how much each descriptor decreases the degree of impurity in the corresponding tree: the more a variable decreases the error at the split node, the more the variable is important. In particular, the predictor importance measure is calculated for each tree, then averaged over the total RF, and, finally, divided by the STD over the whole ensemble algorithm. The implemented FS method leads to the final selection of all those radiomic features whose importance exceeds a threshold set at both 0.05 and 0.1 (corresponding to the significance levels of 5% and 10%).

3.4.3 Dataset Balancing

Another way to improve the classification performance of the ML model is to balance the number of elements from different classes (benign and malignant in this study) that belong to the training set. Indeed, what frequently happens in classification problems is to find a high dataset imbalance, especially in the case of healthcare-related applications where at least one of the classes consists of fewer samples: actually, medical imaging datasets present a marked data disproportion between the two classes, with a number of positive cases lower than the negative ones. Taking into consideration the dataset of this study, the benign to malignant ratio is 3:1 and influences the classification performance, both by creating a bias towards the majority class of benign lesions and compromising the use of some evaluation metrics, which cannot be considered indicative of the comparison between different ML models.

There are several techniques for solving the imbalance dataset problem, which, in most cases, belong to the two categories of *oversampling* and *undersampling* [93]. The former is configured as a replication process that wants to increase the number of elements of the minority class (malignant in this study), for example by generating synthetic data that try to simulate their attributes. The latter is devised

as a process that wants to reduce the number of samples of the majority class (benign), randomly eliminating some of the observations to match its size with the minority class. These balancing process are only applied to the training sets, used in the learning phase of the model. This study implemented the *undersampling* method. In practice, a random selection of a subset of samples of the majority class was performed. Indeed, Breast CT voxels are isotropic, and there are no preferential views: all nine views from each tumor have the same spatial resolution. Considering the three different size datasets, three selection strategies were performed: the *Coronal Plane* training set involved the same number of benign and malignant masses, the *Anatomical Planes* training set was balanced by extracting one patch for each benign mass, while the *9 Planes* training set observed the extraction of three patches for each lesion of the majority class. The final balanced training sets are reported in Table 4.

The undersampling balancing process certainly involves a loss of information that could cause a less generalization on the test set or new datasets, but, at the same time, is useful for the classification both in terms of performance and metrics adequate for the comparison of the developed models. In particular, one of the commonly used metrics is *accuracy*, which is not appropriate with imbalanced datasets but becomes relevant after the class balancing. A detailed description of the measures adopted in this study for the comparison of the different ML models is presented in the following section.

Table 4. Results of the balancing of the three image training sets, with the number of benign and malignant patches assigned to each of them.

Balanced Training Set	Benign	Malignant	Total
Coronal Plane	11	11	22
Anatomical Planes (Coronal, Sagittal, Axial)	36	33	69
9 Planes	108	99	207

3.5 Artificial Neural Network Classification Model

In the field of supervised learning for classification and regression tasks, one of the most successful and widely used algorithms is Artificial Neural Network (ANN), simply called Neural Network (NN). Indeed, nowadays, many classification models adopt NNs for their flexibility and the results achieved. The first subsection provides the NN model representation, describing in detail its architecture (components and organization) and its learning mechanism (training function and performance analysis), and defining the hyperparameters adopted in this thesis. The second subsection focuses on the classification performance metrics used to compare the different NN models implemented during the study and the determination of the best radiomic-based classification model for the discrimination of benign and malignant breast masses imaged with Breast CT.

3.5.1 Model Representation

NNs are a computational model inspired by the functioning of brain neural networks. These systems can learn to perform a task (classification in this study), starting from the input data during the learning phase without having to be programmed with ad-hoc rules. Therefore, the information contained within the training set is used to train the NN, which looks for the underlying patterns between the attributes of the input samples and their respective (known) outcome. NN's structure and learning method are described in the following subsections.

Architecture

The basic unit of an NN is the Artificial Neuron, also called Perceptron, which takes the inputs (like the *dendrites* of the biological correspondent), processes them, and transmits an output (like the *axon*). Indeed, each neuron has an activation function, which maps the appropriately weighted inputs to its output (Figure 31). There are different types of activation functions, including the Heaviside step function (or unit step function), the linear function, and the sigmoid function. Each of them has its pros and cons on the stability of the NN, composed of mutually connected neurons. The activation function adopted in this study for internal neurons (hidden neurons, as will be described below) is the hyperbolic tangent sigmoid function (equation 36), whose range is [-1,1], while the last output neuron presents the logistic sigmoid function (equation 37), whose range is [0,1].

$$\mathrm{TanSig(n)} = \frac{2}{(\,1 + e^{-2 \cdot n}\,)} - 1 \tag{36}$$

$$\mathrm{LogSig(n)} = \frac{1}{(\,1 + e^{-n}\,)} \tag{37}$$

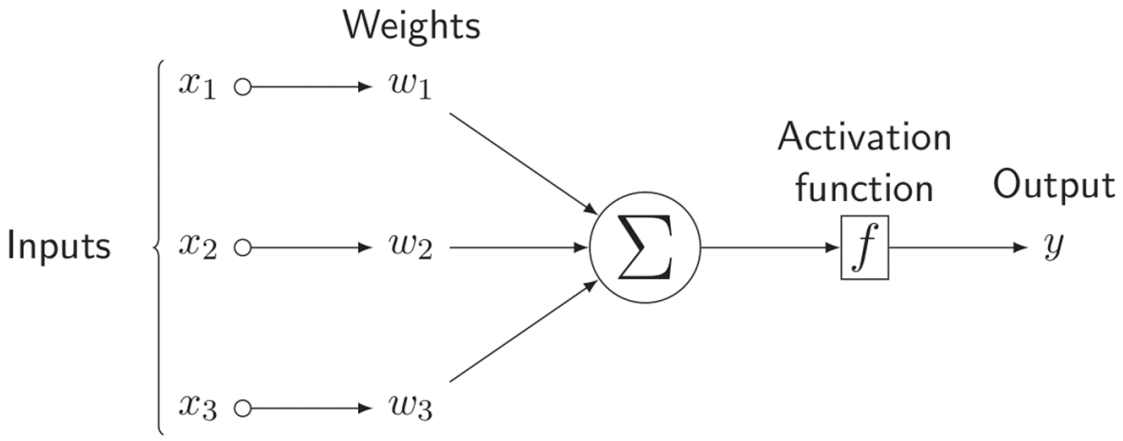


Figure 31. Representation of an Artificial Neuron (Perceptron). Each input variable (x_i) is weighted connected to the neuron, and its incoming contribution is summed together with the others. Then, the neuron applies the activation function to this sum to obtain the output (y), which is passed to all the neurons of the next layer. Retrieved from [94].

From an architectural point of view, NNs typically consist of three types of layers (Figure 32), which can be described as:

- **Input**, which has the task of receiving data directly from the outside world. In our study, the inputs are the values of the features extracted from all the patches that compose three sets of training, validation, and testing.

- **Output**, which provides the NN prediction regarding the input element, whether in the training or classification phase.

- **Hidden** (optional), which stands between the two previously mentioned layers and is used to calculate more complex hypotheses than a single unit with only input and output. One or more layers are usually included in a NN. They are defined as *hidden* because they are in the middle and can communicate only with other neurons (no direct contact with the outside world). The number of hidden layers and the number of hidden neurons constitute the study's hyperparameters, where a *hyperparameter* is defined as the parameter tuned before the start of the learning process and analyzed based on NN performance to understand which setting has the best results.

Therefore, NN consists of the mutual connection of the neurons that belong to these layers. Indeed, each neuron is interconnected to all the neurons of the next layer via weights, considering that its output becomes their input. The NNs that present this type of architecture are called *feedforward* because the information propagation occurs only forward, without any cycle or loop that can carry information to neurons belonging to the same layer.

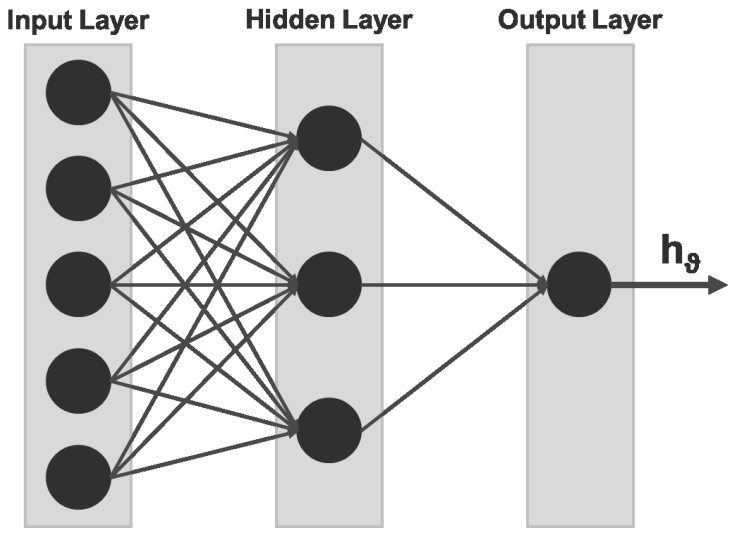


Figure 32. Representation of a typical NN's architecture. There are an input layer, an output layer, and one or more hidden layers. The input layer receives the input data from the outside, the output layer provides the NN's prediction, and the hidden layers connect these two layers and create more interconnections between the neurons of the NN.

Learning Mechanism

Learning is the process performed before using the NN with the testing set or new samples to make the model fit for the classification task for which it is designed. The learning method widely used in the feedforward NNs is the *Backpropagation*, a minimization algorithm that computes the set of optimal parameters relative to the weights of the connections. The goal is to adjust these weights to minimize the error that occurs in the prediction of training samples. In particular, this error estimation process is performed by iterating in the opposite direction to that of information propagation and takes the name of backpropagation (Figure 33). From an operating point of view, this learning algorithm efficiently calculates the gradient associated with the cost function, trying to minimize it by searching for the best set of weights that connect the various NN layers. In this study, the *scaled conjugate gradient backpropagation* was adopted, which proved to be faster but also more sensitive to the random initialization of the NN weights than the traditional backpropagation algorithm [95].

The learning process requires a training set because the input elements are those used to calculate the error. This process, repeated for all the training set samples, can be described in the following steps:

1. Initial feedforward propagation for the output prediction of the sample.
2. Calculation of the classification error, that is simply the difference between the result obtained and the real output. In this study, the performance

analysis was done using the *cross-entropy* function, which penalizes the predicted outputs far from their target output.

3. Updating simultaneously the NN weights through the Backpropagation algorithm to reduce the prediction error. The errors that refer to the hidden layers are calculated by multiplying the error corresponding to the next layer with the matrix of the weights relative to the connections between the two layers, and with the activation function derivative evaluated with the input values of the layer being analyzed. Therefore, the process is backward because it starts from the last output layer and returns to the input layer.

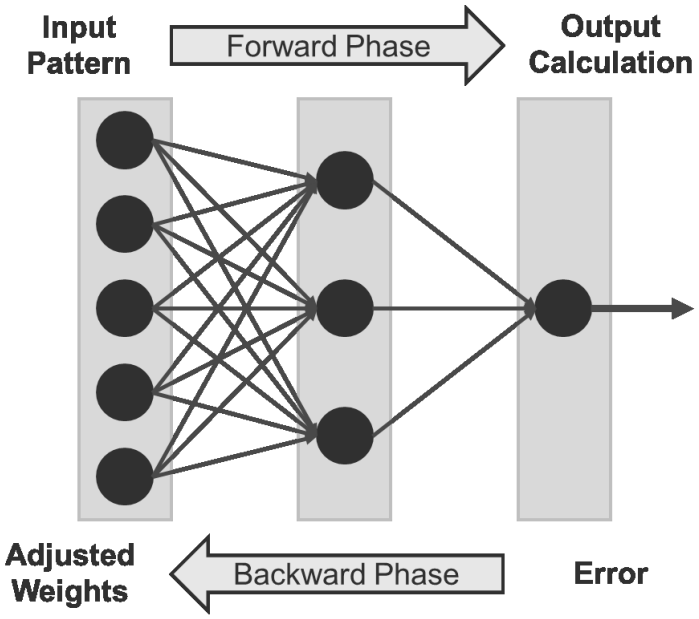


Figure 33. Schematic representation of the Backpropagation learning algorithm. It consists of the forward and backward phases. The former starts from the NN inputs and reaches the output prediction. The latter begins with the error obtained with the classification of the sample (defined as the difference between the prediction and the real outcome) and updates simultaneously the weights of the connections between the neurons that belong to the different layers of the NN.

The appropriate training, which allows avoiding the overfitting problem, and the design of the best architecture, are the main difficulties in the NN's implementation. In this study, the parameters defined for the learning process are listed below:

- Maximum number of epochs reached during training: $\mathbf{1000}$;
- Minimization goal to be achieved (performance goal): $\mathbf{0}$;
- Minimum gradient performance (gradient threshold): $\mathbf{1e^{-6}}$;
- Maximum number of validation failures (the validation set error must increase a fixed number of times after the last time it was decreased, to interrupt the training process): $\mathbf{6}$.

3.5.2 Classification Performance Metrics

At the end of the construction of each NN model, it is necessary to evaluate the performances obtained through metrics that quantify their effectiveness in the classification task. In particular, most of the evaluation metrics are related to the *Confusion Matrix* (CM), which returns a representation of the classification performed in terms of correct and incorrect predictions. Indeed, CM is a matrix where columns refer to the real classes of the classified samples, while rows contain the predictions made by the classifier. Considering a binary classification problem (as faced in this study), CM is configured as a 2x2 matrix, where there are four different outcomes depending on the real class and the prediction of each instance (Figure 34). In particular, each combination can be defined as:

- **True Negative (TN)**, indicating a negative element correctly classified as such (e.g. a benign mass predicted as benign).
- **True Positive (TP)**, indicating a positive element correctly classified as such (e.g. a malignant mass predicted as malignant).
- **False Negative (FN)**, indicating a positive element incorrectly classified as negative.
- **False Positive (FP)**, indicating a negative element incorrectly classified as positive.

		True Class	
		0	1
Predicted Class	0	True Negative (TN)	False Negative (FN)
	1	False Positive (FP)	True Positive (TP)

Figure 34. Representation of a Confusion Matrix for a binary classification problem. The columns refer to the actual classes, while the rows identify the predicted ones. The four combinations define the types of the elements classified correctly (principal diagonal) and incorrectly (antidiagonal).

From the combination of the information extracted from the CM, it is possible to calculate a series of performance measures. In this study, the metrics used to compare the different predictive NN models are 8 and presented below.

Sensitivity, also called **Recall** or **True Positive Rate (TPR)**, is defined as:

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{38}$$

It provides an estimate on the elements belonging to the positive class (malignant tumors), putting the number of correctly classified positive elements in relation to all the samples of this class classified both as positive and negative.

Specificity, also called **True Negative Rate (TNR)**, is defined as:

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{39}$$

It analyzes the elements of the negative class (benign tumors) by making the ratio between the number of correctly classified negative instances and the whole set of the samples of this class.

Accuracy is defined as:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TN} + \text{TN} + \text{FP} + \text{FN}} \tag{40}$$

It defines how accurate the model is, by evaluating the number of times the classifier correctly predicts the negative and positive elements. Considering the CM, this metric is determined as the sum of the elements that are on the principal diagonal divided by the total number of elements classified by the algorithm.

Positive Predictive Value (PPV), also called **Precision**, is defined as:

$$\text{PPV} = \text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{41}$$

It evaluates the classification capability based on the elements predicted as positive. It is measured as the ratio between the number of elements correctly classified as positive and the number of elements for which the model gives a positive prediction.

Negative Predictive Value (NPV) is defined as:

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}} \tag{42}$$

It is an indicator based on the elements predicted as negative and is measured as the ratio between the elements correctly classified as negative and all the elements correctly and incorrectly predicted as negative.

F1-Score is defined as:

$$\text{F1 score} = \frac{2 \cdot \text{PPV} \cdot \text{sensitivity}}{\text{PPV} + \text{sensitivity}} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}} \tag{43}$$

It is also a measure of the classifier accuracy and is obtained from the harmonic average of *sensitivity* and *PPV*. It represents an alternative metric to *accuracy*.

Area Under Curve (AUC) is defined as the area underneath the *Receiver Operating Characteristic* (ROC) curve and provides a measure of the classifier's performance by considering all the possible thresholds set for the classification. Indeed, the threshold of 0.5 is the one typically used, but it can range between 0, where all the elements are classified as positive, and 1, where all are classified as negative. The ROC curve is the graph obtained by taking all the (*1-specificity*, *sensitivity*) pairs calculated at different decision thresholds (Figure 32). Referring to the description of the CM combinations, the ROC curve is drawn by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at several classification thresholds. They can be defined as:

$$\text{TPR} = \text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{44}$$

$$\text{FPR} = 1 - \text{specificity} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{45}$$

The ROC curve is based on elements predicted as positive, which belong to the two separate positive (TP and FN) and negative (FP and TN) classes. The goal is to build a classifier whose curve is in the upper-left corner, where ideally the maximum sensitivity and specificity are achieved. Therefore, AUC is a measure that allows understanding how well a classification model predicts, ranging from 0 to 1.
In particular, if the AUC is around 0.5, then the classifier is not considered reliable because it works randomly. If the AUC is 1, then the classifier is perfect and the predictions are 100% correct. This metric is threshold-invariant, measuring the prediction quality regardless of the decision threshold. However, this aspect is often not useful because it can be significant to assess which is the trade-off threshold for having specific performance, such as the increase of sensitivity at the expense of specificity. It is not the best parameter if an optimization process for choosing the best decision threshold is carried out. Moreover, AUC can be evaluated from the graph that considers the (*sensitivity*, *PPV*) pairs, which is called the *Precision-Recall*

(PR) curve. This curve examines only the TPs (indeed, the TNs do not appear in either of the two parameters), and the model is supposed to be good if the PR curve is in the upper right corner, where ideally the maximum sensitivity and precision are achieved (Figure 35). Although the AUC related to the ROC Curve is the most known and used, this PR metric has its relevance, and, like the other AUC, the higher the value, the better the model is. To distinguish them, the two AUC values were called **AUC_{ROC}** and **AUC_{PR}** to refer to the one calculated from the ROC and the PR curves, respectively.

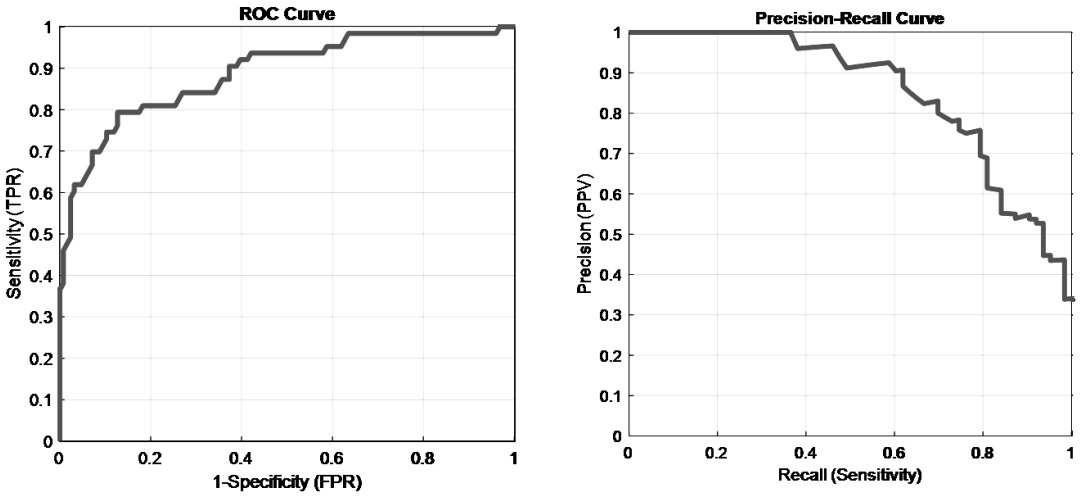


Figure 35. Examples of Receiver Operating Characteristic (ROC) and a Precision-Recall (PR) curves of an NN classification model developed in this study. These two curves can be considered relatively good because they are in the upper-left and upper-right corners, respectively. Indeed, AUC_{ROC} and AUC_{PR} are 0.89 and 0.84, respectively.

Actually, these performance metrics are not suitable for being applied at every stage of the model-building process. Indeed, some measures cannot be assessed with imbalanced datasets because their value does not take this aspect into account. *Accuracy* is one of them since it considers all the elements correctly classified, and is not applicable if there is a disproportion between the majority and minority classes because it could have a high value even making many errors on the minority class. At the same time, some metrics are particularly useful, depending on the goal to be achieved. In this study, the purpose is to evaluate the model's prediction capability on the minority class (malignant tumors) compared to the majority class, and *F1-Score* was chosen as the most suitable metric to define the best models in all the different development phases, even adopting the initial imbalanced datasets. Furthermore, *sensitivity* cooperated to give an overall view of the NN model analysis. Even when the datasets were balanced, *F1-Score* and *sensitivity* had more importance, as a matter of continuity during the entire study, but also the other measures were taken into consideration in defining the classifier performance. Referring to the comparison between the various Feature Selection methods, the AUC_{ROC} and AUC_{PR} were used for a comprehensive evaluation of their performance.

4 Results

In this chapter, all relevant results obtained during the design of the Neural Network (NN) classification models are shown. The training, implementation, and evaluation processes carried out to determine the best NN model can be divided into three different analyses, corresponding to the first three sections of this chapter.
In particular, the first study (Section 4.1) was conducted by directly evaluating the initial dataset without any feature processing or training set modifications. The second study (Section 4.2) involved the analysis of the different Feature Selection (FS) methods and a comparison between the classification performance before and after FS. The third study (Section 4.3) focused on the dataset balancing process to understand its classification benefits, evaluating both the context without and with FS and making a comparison with the previously developed NN classifiers.
Finally, Section 4.4 describes the final NN model and a detailed discussion on the contribution of the two procedures performed on the datasets in terms of features (*Feature Selection*) and samples (*Dataset Balancing*).
At first, all three different size datasets were taken into account to highlight the difficulties in adopting limited datasets for the development of classification models and, therefore, the relevance of a dataset consisting of a number of elements that is higher than the number of features. Subsequently, only the dataset containing the 9 image patches for each mass (*9 Planes*) was chosen for the subsequent analyses.

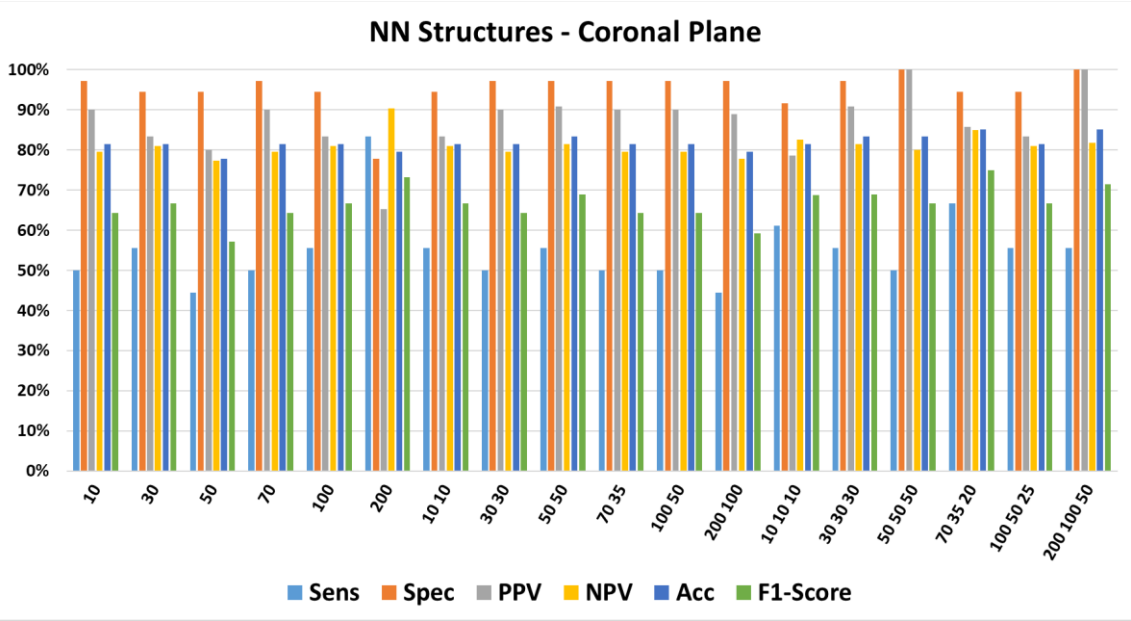
4.1 1st Analysis: Original Dataset

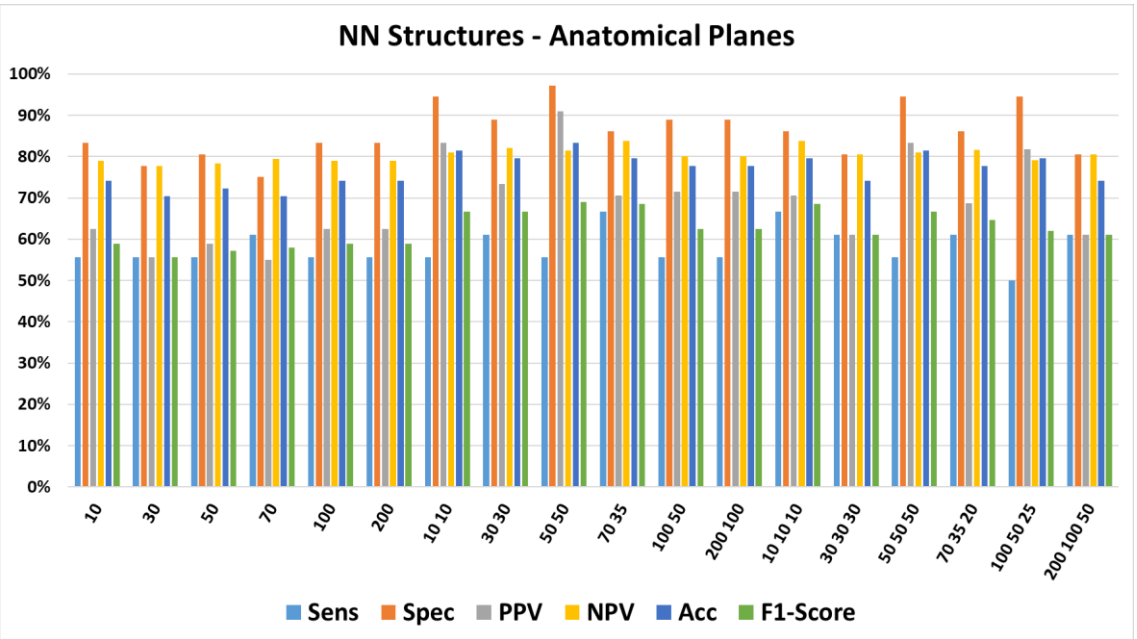
The first analysis involved the three original and imbalanced image datasets, appropriately divided into the three sets of training, validation, and test (Table 2). All radiomic descriptors were used, a total of 158 attributes characterizing each sample of the datasets.
The NN architecture study aimed to find the right combination between the number of hidden layers and the number of neurons belonging to each of them. Indeed, there are no studies on medical image pattern recognition that indicate which are the best NN structures to implement, since each classification problem requires the fine tuning of the network architecture according to the training set size and distribution. Thus, one-, two-, and three-layer NNs with a number of neurons per hidden layer ranging between 10 and 200 were implemented (Table 5). The choice to have at most 3 hidden layers and 200 neurons per level was made so as to avoid to build excessively complex hidden layers that could also have increased in the computational cost to train each implemented NN model (and could increase overfitting).

Table 5. Overview of the different NN architectures that were investigated in this study. The one, two, and three hidden layers structures that were implemented are shown in the three columns with the number of hidden neurons included.

Single Layer	Double Layer		Triple Layer		
10	10	10	10	10	10
30	30	30	30	30	30
50	50	50	50	50	50
70	70	35	70	35	20
100	100	50	100	50	25
200	200	100	200	100	50

As regards the training process, the process used each training set, internally divided into 90% training and 10% validation, while the testing was carried out on the validation set of the *9 Planes (9P)* dataset, since it consisted of the largest number of samples (54 patches). This choice was made not to bias the classifier on the test set: indeed, the architecture hyperparameters were tuned starting from the NN performance on the validation set and were not changed in the final testing phase. Whenever an NN is trained with the approach followed in this thesis, different classification results can be obtained due to the random weight initializations and division of the training set in training and validation subsets. Consequently, 10 iterations for each of the 18 NN structures were performed, choosing the best model for each dataset based on the highest *F1-Score* value (Figure 36).





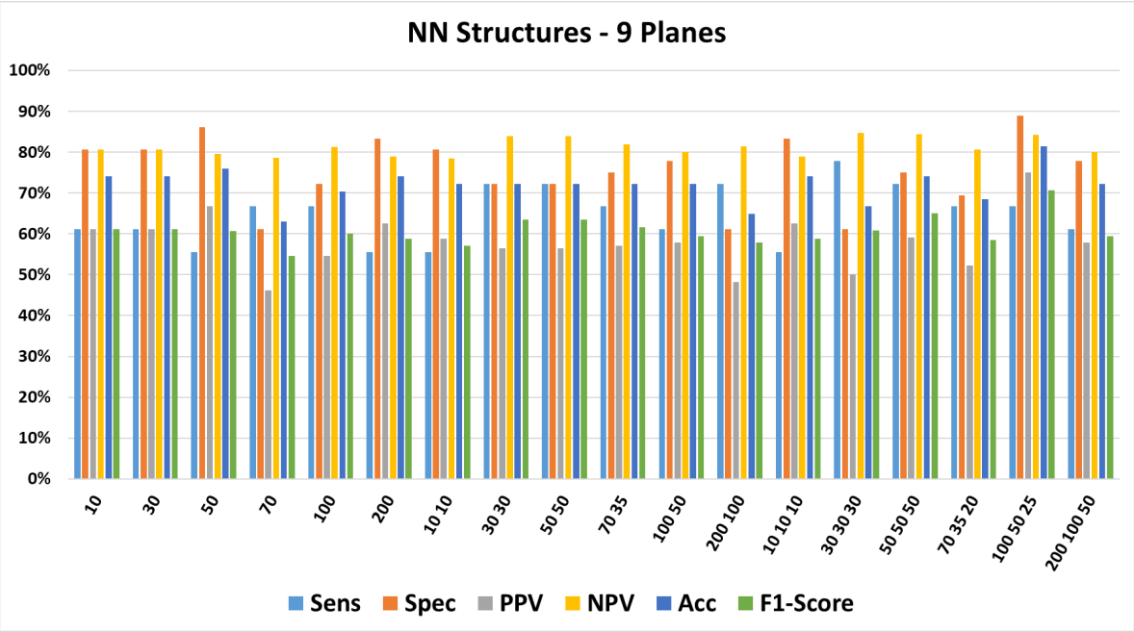


Figure 36. Validation results of the 18 models built with a different architecture for each of the three original datasets (*Coronal Plane*, *Anatomical Planes*, *9 Planes*). For each structure, the values of 6 metrics (*sensitivity*, *specificity*, *PPV*, *NPV*, *accuracy*, *F1-Score*) referring to the best of the 10 iterations are illustrated. The choice of the best architecture was based on the *F1-Score* (last bar, in green): in particular, the three best NN structures were [70 35 20], [10 10 10] and [100 50 25] for the *Coronal Plane*, *Anatomical Planes*, and *9 Planes* datasets, respectively.

The best structures for each of the three datasets were evaluated on the test set of the *9P* dataset, since it had the largest number of samples (189 Patches). The CMs and the metric values of the three classifiers are shown in Figure 37. The NN model with the best *F1-Score* performance was identified as the "winning" model. In case of same *F1-Score* values, the *sensitivity* was used to define the best NN model.

CP		Target	
		0	1
Predicted	0	124	43
	1	2	20

AP		Target	
		0	1
Predicted	0	122	38
	1	4	25

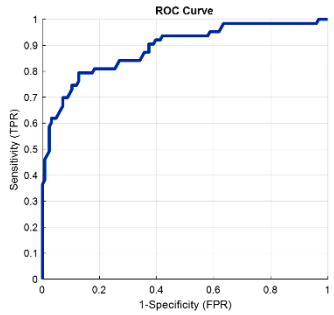
9P		Target	
		0	1
Predicted	0	124	34
	1	2	29

Image Dataset	Structure	Sens	Spec	PPV	NPV	Acc	AUC_{ROC}	AUC_{PR}	F1-Score
Coronal Plane	CP [70 35 20]	32%	98%	91%	74%	78%	0.76	0.68	47%
Anatomical Planes	AP [10 10 10]	40%	97%	86%	76%	3%	0.83	0.76	54%
9 Planes	9P [100 50 25]	46%	98%	93%	78%	81%	0.79	0.76	62%

Figure 37. Performance representation of the best NN models associated with each of the three original datasets. The Confusion Matrices (CMs) obtained from the classification of the *9 Planes* test samples, and the performance indicators are reported for each model. The NN classifier based on the *9 Planes* dataset with a [100 50 25] architecture was the best of this comparison because it had the highest *F1-Score* value (last column of the table, in blue).

As can be seen from the results, the classifier model obtained using the *Coronal Plane (CP)* dataset is overfitted on the training set and is not able to generalize on the new test samples, obtaining an *F1-Score* and a *sensitivity* of 47 % and 32%, respectively. The best NN model of the *Anatomical Planes (AP)* dataset was better with 54% of *F1-Score* and 40% of *sensitivity*, even if it presents a high ratio between the number of attributes and number of samples too. Finally the best NN model of the *9P* dataset achieved the best performance of *F1-Score* (62%) and *sensitivity* (46%), considering the higher number of training elements that provided a better representation of the breast mass population and exceeded the number of extracted radiomic descriptors (the number of features was about 2.5 times lower than the number of *9P* training samples). Therefore, the *9P* model with its [100 50 25] hidden-layer architecture was kept and not modified further for this first part of the study. In particular, a new training was performed by using both the training and validation set for the training phase, and the test set for assessing its generalization performance. This process was repeated 100 times because the random initialization of the weights could undermine the classification results [Section 3.5]. The classifier among the 100 developed with the highest *F1-Score* value was chosen as the final NN model of this analysis with the use of original datasets without any data processing. The results (CM, metric values, ROC curve, and PR curve) relating to this NN classifier are shown in figure 38.

9P		Target	
		0	1
Predicted	0	122	26
	1	4	37



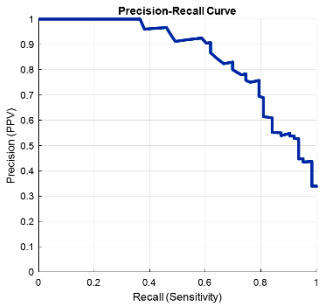


Image Dataset	Structure	Sens	Spec	PPV	NPV	Acc	AUC_{ROC}	AUC_{PR}	F1-Score
9 Planes	**9P [100 50 25]**	59%	97%	90%	82%	84%	0.89	0.84	**71%**

Figure 38. Illustration of the characteristics of the best NN classification model starting from the original datasets without any data processing (no Feature Selection or Dataset Balancing). This model was obtained at the end of the retraining process, where 100 training iterations were performed keeping the hyperparameters (number of hidden layers and neurons) fixed. It showed the best generalization performances on the *9 Planes* test set in terms of *F1-Score* and *sensitivity*.

The *9P* dataset (as well as the other two smaller datasets) in its initial condition was not equally represented by the two tumor groups and showed a notable imbalance towards the class of benign lesions (3 to 1 compared to malignant ones), which results in the model's abilities to classify correctly these lesions (*specificity* of 97% for the final NN model), as it is biased towards the most frequently represented class (i.e. benign). In fact, the results obtained so far in terms of *F1-Score* and *sensitivity* are not satisfactory, and the goal is to ensure that they can be improved by considering some modifications in the number of features or training samples, such as the FS application analyzed in the next section.

4.2 2nd Analysis: Feature Selection

The second analysis concerned the application of the 9 FS methods described in Subsection 3.4.2. The dataset used is that of the *9 Planes (9P)*, on which the NN model obtained at the end of the first analysis is based (Figure 38). The objective is to understand if it is possible to improve classification performance by reducing the number of attributes to be considered in the NN training process. Indeed, as previously described, FS removes all those variables that prove to be uninformative, irrelevant, and redundant for the task to be performed. From an implementation point of view, for each FS method, the feature extraction process on the *9P* dataset was first performed and, subsequently, the training phase was carried out for 100 iterations, adopting the training and validation sets consisting solely of the features considered relevant. The number of features extracted from each FS strategy using the initial *9P* dataset is shown in Table 6.

Table 6. Overview of the different Feature Extraction methods investigated in this study. This table shows the number of features extracted from each strategy starting from the original (and imbalanced) *9 Planes* dataset.

FS Method	# of Features Extracted
Correlation 0.90	75
Correlation 0.80	59
Mann-Whitney U-Test	110
Mann-Whitney U-Test & Correlation 0.90	80
Decision Tree	10
Random Forest (CART & 0.05)	79
Random Forest (CART & 0.10)	47
Random Forest (Curvature & 0.05)	101
Random Forest (Curvature & 0.10)	68

Similarly to the first analysis (Section 4.1), the generalization performances of the 100 models were evaluated on the *9P* test set. The model among them with the highest *F1-Score* achieved was chosen to represent the FS method in the comparison made with the other ones to understand which is the best. The results relating to the different FS techniques based on the performance metrics are shown in Table 7.

Table 7. Performance evaluation of the NN models associated with each Feature Selection method. All the models had a [100 50 25] architecture, and were trained using the 9P training and validation sets. The metrics shown in the table were obtained from the classification of the *9 Planes* test samples. The NN classifier model built with the 75 features extracted using the *Correlation-based (threshold of 0.90)* FS method presented the highest *F1-Score* value (last column).

FS Method	Sens	Spec	PPV	NPV	Acc	AUC_{ROC}	AUC_{PR}	F1S
Correlation 0.90	62%	96%	89%	83%	85%	0.85	0.80	**73%**
Correlation 0.80	60%	94%	83%	82%	82%	0.83	0.79	**70%**
Mann-Whitney U-Test	56%	98%	95%	82%	84%	0.85	0.80	**70%**
Mann-Whitney U-Test & Correlation 0.90	59%	98%	92%	82%	85%	0.85	0.81	**72%**
Decision Tree	52%	100%	100%	81%	84%	0.82	0.79	**69%**
Random Forest (CART & 0.05)	57%	98%	92%	82%	84%	0.85	0.81	**71%**
Random Forest (CART & 0.10)	63%	90%	75%	83%	81%	0.86	0.78	**69%**
Random Forest (Curvature & 0.05)	57%	96%	88%	82%	83%	0.86	0.81	**69%**
Random Forest (Curvature & 0.10)	59%	98%	95%	83%	85%	0.86	0.81	**72%**

The FS strategy that obtained the highest *F1-Score* performance starting from the initial *9P* dataset and with a [100 50 25] NN architecture was the **Correlation with a threshold of 0.90**, which extracted a total of **75 features**.

Actually, by further analyzing the *AUC* values of the ROC and PR curves, it was observed that the performances obtained with different FS methods were all comparable, and there were no significant dissimilarities. Both curves were analyzed because they give complementary information on imbalanced datasets. Indeed, the ROC curve, used as a comparison metric in most research studies, refers to *sensitivity* and *specificity*, and, therefore, can be affected by skewness in the class distribution. Instead, the PR curve evaluates two ratios associated with TPs (i.e. *sensitivity* and *precision*) and can provide an indication that goes beyond the class imbalance [96]. These results that represent this absence of difference between the various FS methods are illustrated in Figure 39.

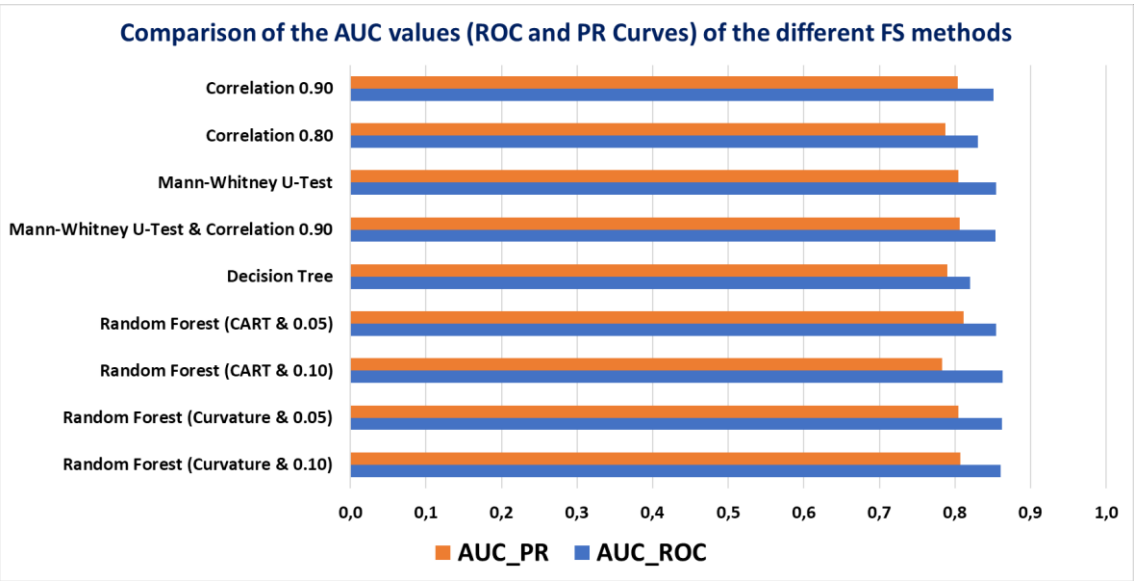
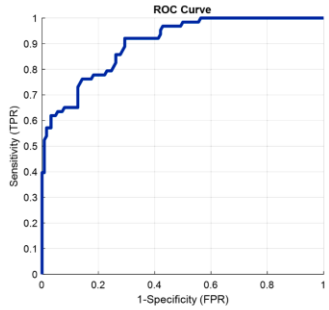


Figure 39. Bar chart for comparing all FS methods based on the AUC values of the ROC (*AUC_ROC*) and PR (*AUC_PR*) curves. It shows that the influence of the choice of the FS strategy on classification performance was not significant since AUCs were equivalent between all the FS methods. Indeed, the AUC_{ROC} values ranged between 0.82 and 0.86, while the AUC_{PR} ones between 0.78 and 0.81.

After choosing the FS method that provided the best performance scores (although not significantly different from the other techniques), the NN architecture study was done again because the hyperparameters depend on the number of features and training samples. In particular, it is plausible that the most proper NN structure for the condition with a lower number of attributes (and, thus, a lower ratio between features and samples) is smaller and less complex. Therefore, the same investigation performed in the first part of the first analysis (described in Section 4.1) was repeated starting from the *9P* dataset, which consisted of the same number of initial samples and only the 75 features extracted using the *Correlation-based* FS method. The NN classification model with the highest *F1-Score* contained three hidden layers of 10 neurons each. Following the same scheme, the model was retrained on the training and validation samples using the selected set of hyperparameters, and its generalization performances were evaluated on the test set. The results (CM, metric values, ROC curve, and PR curve) of the best iteration are shown in Figure 40.

9P		Target 0	Target 1
Predicted	0	120	24
	1	6	39



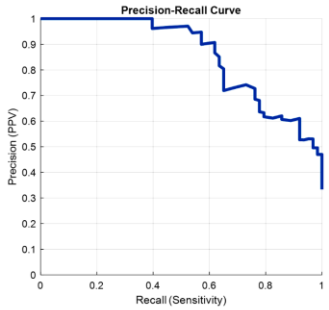


Image Dataset	Structure	Sens	Spec	PPV	NPV	Acc	AUC_{ROC}	AUC_{PR}	F1-Score
9 Planes 9P	[10 10 10]	62%	95%	87%	83%	84%	0.90	0.84	72%

Figure 40. Illustration of the characteristics of the best NN classification model starting from the original *9 Planes* dataset with the application of the *Correlation-based (threshold of 0.90)* FS method. This model was obtained at the end of the retraining stage, where 100 training iterations were performed maintaining the same [10 10 10] structure. It exhibited the highest *F1-Score* value on the test set.

At this stage in the study, it is interesting to compare the various winning models achieved, deepening all the complete analysis in the last Section of this Chapter.
In particular, from the comparison between the model before and after the FS, it can be noted that there was no significant improvements in *F1-Score* and *sensitivity* performance (Figure 41). Despite this, given the slight improvement made with the *Correlation-based* FS, the final NN classifier model of this second analysis became the best model implemented so far, considering the original *9P* dataset imbalanced towards the benign class. The next activity to be performed to improve the classification performance was, actually, the dataset balancing, which could lead to a more evident improvement than that obtained by the application of the FS alone.

PRE		Target 0	Target 1
Predicted	0	122	26
	1	4	37

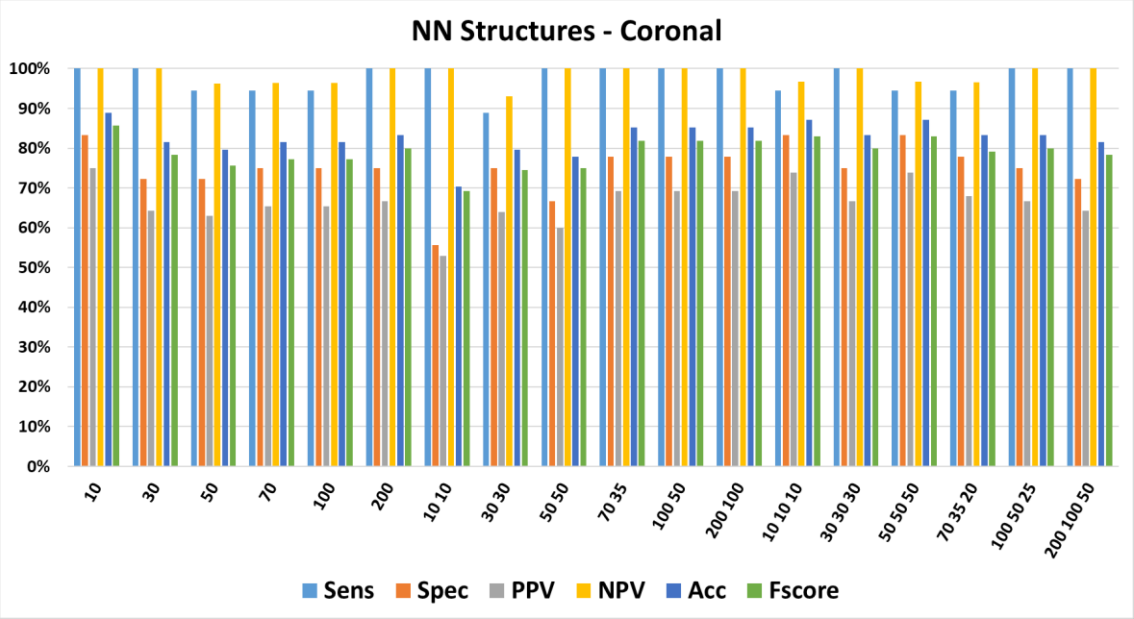
POST		Target 0	Target 1
Predicted	0	120	24
	1	6	39

NN Model	Structure	Sens	Spec	PPV	NPV	Acc	AUC_{ROC}	AUC_{PR}	F1-Score
Before FS PRE	[100 50 25]	59%	97%	90%	82%	84%	0.89	0.84	71%
After FS POST	[10 10 10]	62%	95%	87%	83%	84%	0.90	0.84	72%

Figure 41. Performance representation of the best NN models obtained before (PRE) and after (POST) the application of the *Correlation-based (threshold of 0.90)* FS method. The CMs obtained from the classification of the *9 Planes* test samples, and the performance indicators are reported for each model. The NN classifier obtained after the FS was slightly better (higher *F1-Score*) and was chosen as the best model at the end of these first two analysis stages.

4.3 3rd Analysis: Dataset Balancing

The third analysis focused on the evaluation of the impact that dataset balancing could have on the classification performance. As reported in Subsection 3.4.3, an *undersampling* method was adopted to reduce the number of benign samples and to train the NN models with a similar number of elements representing the two tumor classes. Therefore, starting from the three balanced training sets (Table 4) and considering all the 158 variables initially extracted, the NN architecture study was carried out by implementing the structures defined by the combinations of hidden layers and hidden neurons described in Table 5. The characteristics of the learning process were the same as those implemented in Section 4.1. Indeed, the training set was divided into 90% training and 10% validation for the learning, and the validation set was used for the testing in the building process of 18 models for each dataset. The best iteration of the 10 performed for each of the 18 structures (highest *F1-Score*) was chosen, and all their performance metrics are shown in Figure 42.



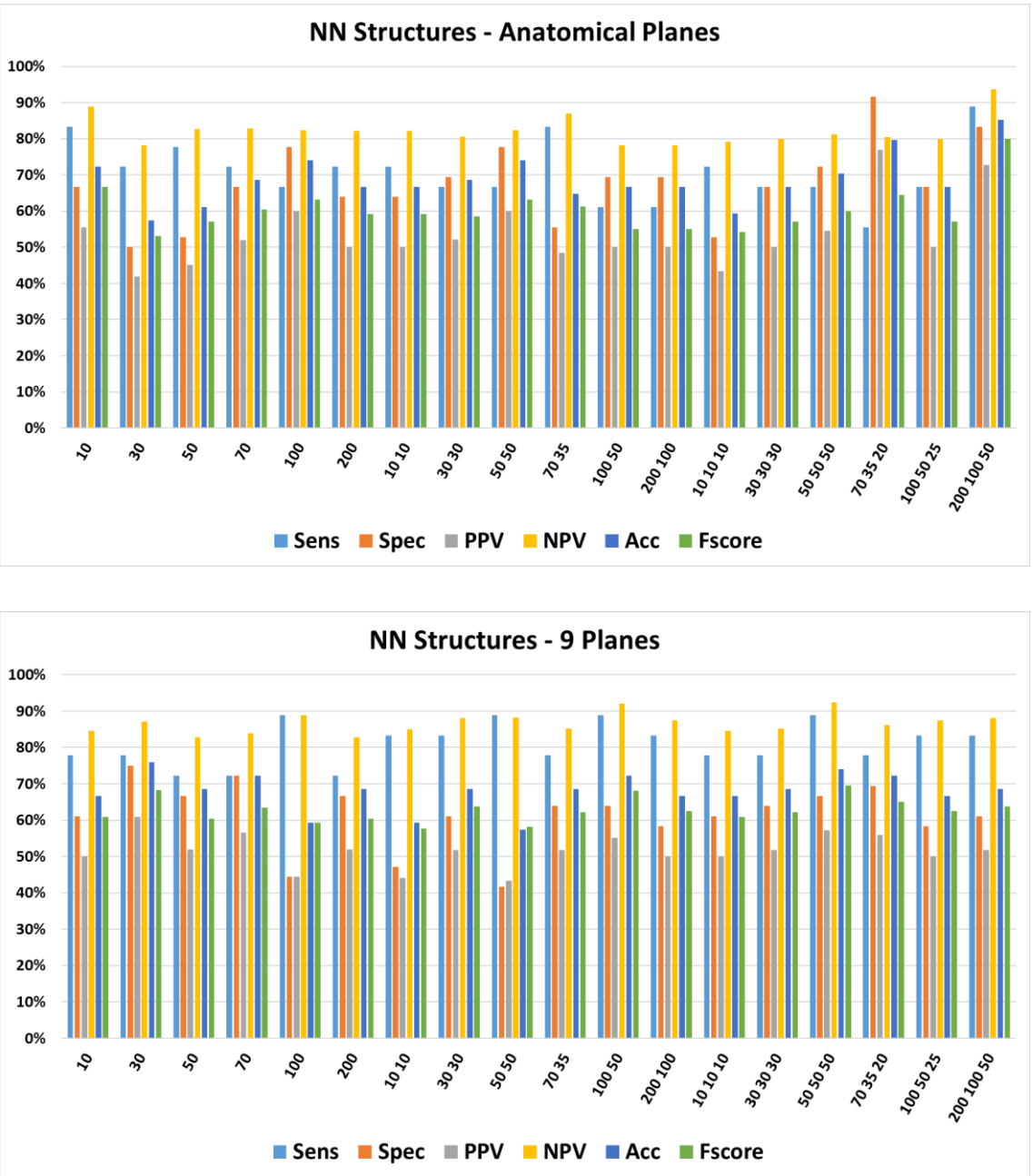


Figure 42. Validation results of the 18 models built with a different architecture for each of the three balanced datasets (*Coronal Plane*, *Anatomical Planes*, *9 Planes*). For each structure, the values of 6 metrics (*sensitivity*, *specificity*, *PPV*, *NPV*, *accuracy*, *F1-Score*) referring to the best of the 10 iterations are illustrated. The choice of the best architecture was based on the *F1-Score* (last bar, in green): in particular, the three best NN structures were [10], [70 35 20] and [50 50 50] for the *Coronal Plane*, *Anatomical Planes*, and *9 Planes* datasets, respectively.

The best three NN models built on each of the three image datasets were evaluated in their generalization abilities on the *9 Planes* test set. The results (CMs and performance metrics) are illustrated in Figure 43.

CP		Target	
		0	1
Predicted	0	106	28
	1	20	35

AP		Target	
		0	1
Predicted	0	109	24
	1	17	39

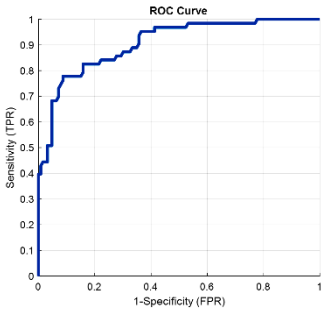
9P		Target	
		0	1
Predicted	0	102	16
	1	24	47

Image Dataset		Structure	Sens	Spec	PPV	NPV	Acc	AUC_{ROC}	AUC_{PR}	F1-Score
Coronal Plane	CP	[10]	56%	84%	64%	79%	75%	0.74	0.68	59%
Anatomical Planes	AP	[70 35 20]	62%	86%	70%	82%	78%	0.80	0.74	65%
9 Planes	9P	[50 50 50]	75%	81%	66%	86%	79%	0.88	0.81	70%

Figure 43. Performance representation of the best NN models associated with each of the three balanced datasets (without Feature Selection). The Confusion Matrices (CMs) obtained from the classification of the *9 Planes* test samples, and the performance metrics are reported for each classification model. The NN classifier based on the *9 Planes* dataset with a [50 50 50] architecture had the highest *F1-Score* value and was chosen as the best model of this comparison.

Even in this analysis, as expected, the best NN classifier was the one obtained starting from the largest dataset (*9 Planes*) with an *F1-Score* of 70%, against 59% and 65% of the *Coronal Plane* and *Anatomical planes* datasets, respectively. This model contained three hidden layers (50 neurons each) and was selected for the second part of the model construction, where, keeping its architecture fixed, it was retrained 100 times (the training and validation sets used for training, the test set for generalization evaluation and comparison). The best NN classification model developed in these 100 iterations was the one that presented the highest *F1-Score* value, and its classification performances (CM, metrics, ROC curve, and PR curve) are shown in figure 44.

9P		Target	
		0	1
Predicted	0	111	14
	1	15	49



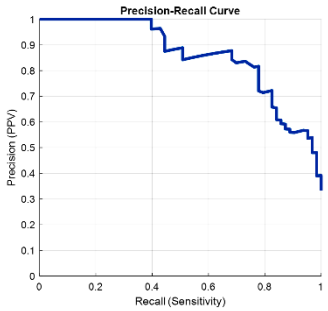


Image Dataset		Structure	Sens	Spec	PPV	NPV	Acc	AUC_{ROC}	AUC_{PR}	F1-Score
9 Planes	9P	[50 50 50]	78%	88%	77%	89%	85%	0.91	0.84	77%

Figure 44. Illustration of the characteristics of the best NN classification model starting from the balanced *9 Planes* dataset (without Feature Selection). This model was obtained at the end of the retraining stage, where 100 training iterations were performed maintaining the same [50 50 50] structure. It exhibited the highest *F1-Score* value on the test set.

Putting aside the FS process implemented in the second analysis (Section 4.2), it is possible to compare the model just built on the balanced dataset with the model obtained from the original and imbalanced dataset (Section 4.1). The comparative analysis of their performance is shown in Figure 45. In particular, the current NN model with a [50 50 50] architecture proved higher performance both in terms of *F1-Score* (77% compared to 71% of the previous model) and *sensitivity* (78% against 59%). However, this improvement led to a decrease in the classification of benign lesions (*specificity* of 88% compared to 97%), motivated by the fact that the new NN classifier was trained considering a similar number of elements belonging to the two tumor classes. At the same time, it is interesting to note the appreciable improvement in the classification of malignant lesions compared to what was achieved by the FS alone, demonstrating the importance of balancing the datasets.

PRE		Target	
		0	1
Predicted	0	122	26
	1	4	37

POST		Target	
		0	1
Predicted	0	111	14
	1	15	49

NN Model		Structure	Sens	Spec	PPV	NPV	Acc	AUC_{ROC}	AUC_{PR}	F1-Score
Before DB	PRE	[100 50 25]	59%	97%	90%	82%	84%	0.89	0.84	71%
After DB	POST	[50 50 50]	78%	88%	77%	89%	85%	0.91	0.84	77%

Figure 45. Performance representation of the best NN models obtained before (PRE) and after (POST) the Dataset Balancing of the 9 Planes dataset (without Feature Selection). The CMs obtained from the classification of the *9 Planes* test samples, and the performance indicators are reported for each model. The NN classifier obtained after the Dataset Balancing showed higher *F1-Score* and *sensitivity* values, underling the important of having balanced classes for classification tasks.

Similarly to Section 4.2, the 9 FS methods were also applied on the balanced dataset to investigate the further improvement brought by FS. This FS process led to the selection of different radiomic feature subsets with respect to those previously obtained: the number of attributes extracted is shown in Table 8.

Table 8. Overview of the different Feature Extraction methods investigated in this study. This table shows the number of features extracted from each strategy starting from the balanced *9 Planes* dataset.

FS Method	# of Features Extracted
Correlation 0.90	76
Correlation 0.80	59
Mann-Whitney U-Test	102
Mann-Whitney U-Test & Correlation 0.90	81
Decision Tree	8
Random Forest (CART & 0.05)	56
Random Forest (CART & 0.10)	38
Random Forest (Curvature & 0.05)	80
Random Forest (Curvature & 0.10)	51

Considering the *9P* training set with only the features extracted, 100 iterations were carried out for each FS method, and their classification abilities were evaluated on the *9P* test set. The NN models from each FS strategy that showed the highest *F1-Score* were compared, and the best results were achieved by the **Random Forest (Standard CART, threshold of 0.05)** method with a total of **56 features extracted**. Given the intention to understand the FS importance, the AUC values (of ROC and PR curves) obtained from the 9 different analyses were put on the same chart, and pointed out that, even in this case, the choice of the FS method to be used was not critical for the classification performance (Figure 46).

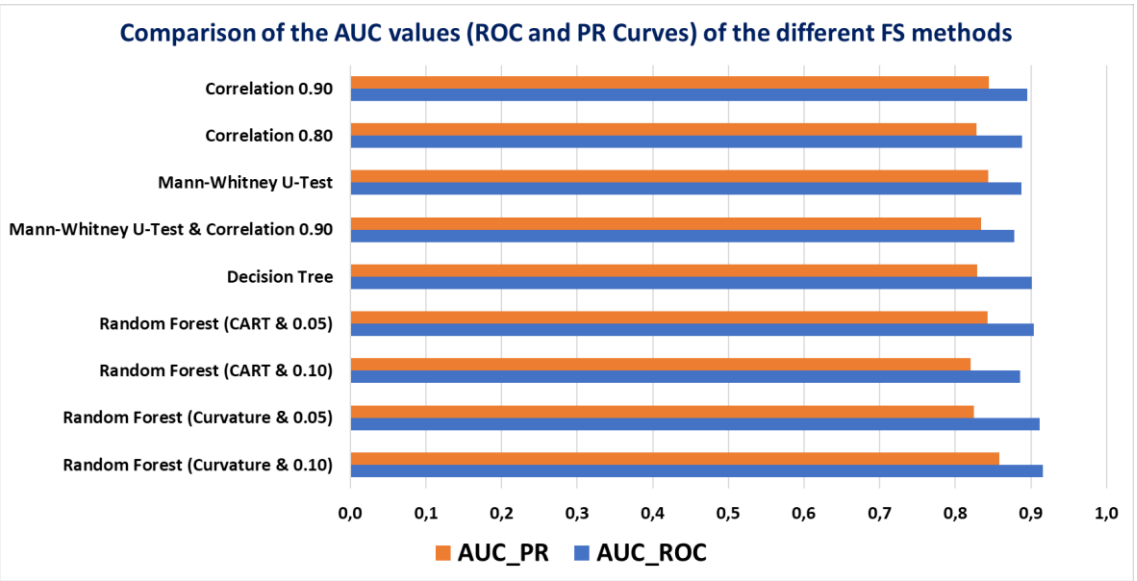
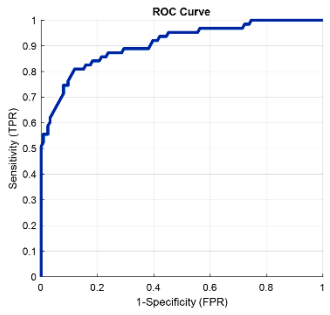


Figure 46. Bar chart for comparing all FS methods based on the AUC values of the ROC (*AUC_ROC*) and PR (*AUC_PR*) curves. Even with balanced datasets, this graph shows that the FS strategy adopted was not important for the significant improvement in classification performance since AUCs were equivalent between all the FS methods. Indeed, the AUC_{ROC} values ranged between 0.88 and 0.92, while the AUC_{PR} ones between 0.81 and 0.85.

After the extraction of the most relevant and informative features, a new study of the NN architecture was carried out to understand if the reduction in the number of attributes brought with it a reduction in the complexity of the classifier model.
At the end of this process, the structure with the best performance consisted of two hidden layers with 10 neurons each, showing a lower complexity than the NN model before the FS. Keeping the structure fixed, 100 retraining iterations were performed, and the best NN model presented generalization performances that reached 79% of *F1-Score* and *Sensitivity*. All the results of this NN model are reported in Figure 47.

9P		Target 0	Target 1
Predicted	0	113	13
	1	13	50



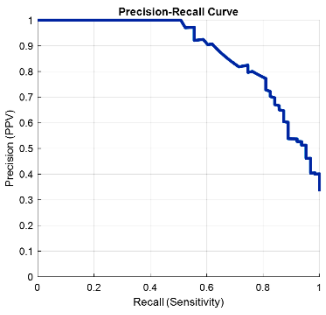


Image Dataset	Structure	Sens	Spec	PPV	NPV	Acc	AUC_{ROC}	AUC_{PR}	F1-Score
9 Planes 9P	[10 10]	79%	90%	79%	90%	86%	0.91	0.86	79%

Figure 47. Illustration of the characteristics of the best NN classification model starting from the balanced *9 Planes* dataset with the application of the *Random Forest (CART Standard, threshold of 0.05)* FS method. This model was obtained at the end of the retraining stage, where 100 training iterations were performed. It exhibited the highest *F1-Score* value on the test set.

To verify what was previously stated regarding the actual importance of the FS, a comparison between the best models trained on a balanced training set before and after the FS was performed. The results showed a slight improvement achieved in terms of *F1-Score*, *sensitivity*, and *accuracy*. In particular, the latter was informative in this analysis since the NN models were built on balanced datasets (Figure 48).

PRE		Target 0	Target 1
Predicted	0	111	14
	1	15	49

POST		Target 0	Target 1
Predicted	0	113	13
	1	13	50

NN Model	Structure	Sens	Spec	PPV	NPV	Acc	AUC_{ROC}	AUC_{PR}	F1-Score
Before FS with DB PRE	**[50 50 50]**	78%	88%	77%	89%	85%	0.91	0.84	**77%**
After FS with DB POST	**[10 10]**	79%	90%	79%	90%	86%	0.91	0.86	**79%**

Figure 48. Performance representation of the best NN models obtained before (PRE) and after (POST) the application of the *Random Forest (Standard CART, threshold of 0.05)* FS method on the balanced *9 Planes* dataset. The CMs obtained from the classification process, and the performance indicators are reported for each model. The NN classifier obtained after the FS was slightly better because it showed higher *F1-Score*, *sensitivity*, and *accuracy*.

On the other hand, the comparison between the last developed NN model (Figure 47) and the model obtained after FS alone (Figure 40) showed the improvement in terms of *F1-Score*, *Sensitivity*, *Accuracy*, AUC_{ROC}, and AUC_{PR} due to the better ability in the classification of malignant tumors, accepting, however, a slight decrease in *specificity*. The results of this analysis are shown in Figure 49. The higher values of these performance metrics obtained thanks to the application of a dataset balancing method, underlined the importance of this process in ML classification problems. Indeed, classes that are equally represented constitute a key element to avoid incurring a bias both in the learning process of the ML model and in the evaluation of specific performance metrics not suitable for imbalanced classes.

PRE		Target 0	Target 1
Predicted	0	120	24
	1	6	39

POST		Target 0	Target 1
Predicted	0	113	13
	1	13	50

NN Model		Structure	Sens	Spec	PPV	NPV	Acc	AUC_{ROC}	AUC_{PR}	F1-Score
Before DB (+FS)	**PRE**	**[10 10 10]**	62%	95%	87%	83%	84%	0.90	0.84	**72%**
After DB (+FS)	**POST**	**[10 10]**	79%	90%	79%	90%	86%	0.91	0.86	**79%**

Figure 49. Performance representation of the best NN classification models obtained before (PRE) and after (POST) the Dataset Balancing and the Feature Selection. The former model was built on the imbalanced *9 Planes* dataset using the 75 features extracted with the *Correlation-based (threshold of 0.90)* FS method, while the latter model was built on the balanced 9 Planes dataset using the 56 features extracted with the *Random Forest (Standard CART, threshold of 0.05)* FS method. The CMs obtained from the classification process, and the performance indicators are reported for each model. The higher values of the performance metrics (*F1-Score*, *sensitivity*, *accuracy*, AUC_{ROC}, AUC_{PR}) of the NN model obtained after balancing the training set showed the importance of having a similar number of samples of the two classes in classification problems.

4.4 Final Neural Network Model and Discussion

The last NN model comparison described in Section 4.3, and illustrated in Figure 49, showed that the model obtained at the end of the whole analysis process, which involved Dataset Balancing and FS, was the one with the highest performance metric values. Its NN architecture contained two hidden layers of 10 neurons each, and a total of 56 relevant and informative features extracted with the *Random Forest (Standard CART, threshold of 0.05)* FS method. In terms of performance metrics, *sensitivity* reached 79%, *F1-Score* 79%, *accuracy* 86%, and AUC_{ROC} 91%, but, at the same time, *specificity* took a relevant value of 90%. For that reason, it constituted the final NN model for the radiomic-based classification of breast tumors of this thesis, and all its performances (CM, metric values, ROC curve, and PR curve) are shown in Figure 47.

A detailed analysis of the several comparisons made throughout Chapter 4 allows drawing some conclusions that concern the two processes investigated to improve the classification performances. In this study, Feature Selection (FS) proved to be a process that, if carried out independently by other operations, did not make a significant addition to improve classification abilities. Indeed, FS was not so critical because the image dataset available was already quite good in terms of the high ratio between the number of examples and the number of attributes, considering the 9 patches for each mass. In particular, Figure 41 shows how the contribution of FS alone with an imbalanced dataset amounted to 1% for *F1-Score* and 3% for *sensitivity*. Likewise, Figure 48 underlines how FS contributed 2% for *F1-Score* and 1% for *sensitivity*, even in the presence of balanced datasets. Therefore, these NN model comparisons led to the conclusion that the choice of the best FS method to be implemented did not prominently influence the classification results, given the small contribution provided in the development of the classification model of breast masses imaged with Breast CT, at least for the (limited) dataset used in this thesis.

On the other hand, Dataset Balancing had a much more significant impact in improving classification performance, especially as regards the malignant tumor class. That is related to the importance of having balanced classes with the same number of elements, which allows NNs to identify all the characteristics of both breast lesion groups and not mainly from the majority one. Figure 45 emphasizes the independent contribution of the Dataset Balancing process, with a 6% increase in *F1-Score* and 19% in *sensitivity*, although accepting a decrease in *specificity* from 97% to 88% due to the reduction in the number of benign elements present in the training set. Finally, the initially mentioned Figure 49 shows the last model comparison that takes into account the cooperation of the two processes, and outlines the characteristics of the final NN classification model, which, in addition to the improvement already mentioned in terms of *sensitivity* (79%), *accuracy* (86%), and *F1-Score* (79%), recovered also in terms of *specificity* (90%).

5 Conclusion

In this thesis, the goal is to design a Computer-Aided Diagnosis (CADx) system for breast tumors classification in Dedicated Breast CT images using shape and margin radiomic features and a Neural Network (NN).
What was developed in terms of descriptors that aim to quantify the morphology and the margin properties of breast tumors, proved to be valuable for the classification task, emphasizing the importance of analyzing different aspects of breast masses rather than only accounting for the traditional and well-known texture features. All these newly developed radiomic descriptors evaluate an aspect of tumoral masses that had never been taken into consideration with this degree of analytical depth and showed how, even if implemented alone, it is possible to achieve classification performance of a high level with the Dedicated Breast CT modality. Naturally, the classification results achieved, although satisfactory in the light of the uniqueness of the implemented descriptors, should be considered preliminary, and underline the need for the continuation of the works to effectively build a CADx system that can provide a second opinion to the radiologist to make a more conscious breast cancer diagnostic decision.

The main limitations of this study mainly concern the small size of the image dataset that was used both in the development phase of the radiomic features and in the training and testing phases of the NN classification models. It was a set of breast lesions that were collected during a clinical trial and proved to be limited, especially for Machine Learning (ML) applications, where numerous training samples are needed. Indeed, all the performances of this thesis were calculated considering each of the 9 patches of the same breast mass as an independent sample. It was done to account for the limited dataset size, and, therefore, provide a larger test set for a stronger performance assessment. The other aspect that needs to be assessed is always related to the dataset but involves the manual annotation that was performed by an image analysis scientist under the supervision of a single expert radiologist. Considering that most of the radiomic descriptors analyze the lesion shape and the irregularities of its contour, the annotation of a single reader could potentially bias the feature analysis and in the classification performance.

Therefore, this study is configured as a starting point for future works that can consider these observations and results, and support the effectiveness and the innovative breadth of the adoption of a radiomic pipeline for classification of benign and malignant breast lesions. Actually, these future studies need larger datasets, which allow implementing CADx systems on a mass-basis (and not patch-basis). By adopting the data augmentation process developed in this thesis, it would be interesting to investigate and implement strategies to merge the classification outcomes from the 9 patches into a single and per-mass diagnostic prediction.

Furthermore, all the breast tumors should be manually segmented by additional experienced radiologists so that the performance of the shape, contour, and margin descriptors can be less biased towards single-expert annotation. A robustness and stability analysis of the proposed radiomic descriptors on different manual segmentations constitutes a study of considerable scope because it could determine an increased involvement of these features in researches and CADx systems.

Besides, it would be necessary to carry out a study that implements a radiomic pipeline that combines both the features developed in this thesis and other types of features (e.g. texture-based features), emphasizing the importance of quantifying all characteristics of the breast lesions with feature sets that contain more than thousands of descriptors.
From an ML point of view, this study implemented only NN classifiers, but there are numerous other supervised algorithms (e.g. Support Vector Machines, Nearest Neighbors, Decision Tree, Random Forest) that have been adopted in other research studies (Chapter 2), and that might provide further insights in classification performance compared to those obtained with this thesis. The evaluation of different ML methods for breast tumor classification is something that should be explored in future works. Likewise, considering the recent advancements of Deep Learning (DL) in many fields of healthcare-related application (including the classification of breast lesions), an interesting aspect is the comparison and synergy of the pipeline of this study based on hand-crafted radiomic descriptors and the DL-based one, where it is not necessary to segment any region of interest or perform any feature extraction and selection.
All these potential and prospective works will contribute to the development of an automated CADx system, which, starting from Dedicated Breast CT images, could predict breast tumors with a higher degree of confidence and accuracy, reducing the number of negative and unnecessary biopsies, which, nowadays, constitutes more than 70% of the overall biopsies performed [14].

Bibliography

[1] World Health Organization International Agency for Research on Cancer (WHO IARC), «Global Cancer Statistics 2018 (GLOBOCAN 2018),» 2019. [Online]. Available: gco.iarc.fr/today/data/factsheets/cancers/20-Breast-fact-sheet.pdf.

[2] World Health Organization, «Breast cancer: prevention and control,» 2019. [Online]. Available: who.int/cancer/detection/breastcancer.

[3] R. L. Siegel, K. D. Miller e A. Jemal, «Cancer statistics 2019,» *CA: A Cancer Journal for Clinicians (American Cancer Society),* vol. 69 (1), pp. 7-34, 2019.

[4] Cancer.net Editorial Board, «Breast Cancer - Statistics,» American Society of Clinical Oncology (ASCO), 2019. [Online]. Available: cancer.net/cancer-types/breast-cancer/statistics.

[5] Associazione Italiana Ricerca sul Cancro (AIRC), "Tumore del seno," 2018. [Online]. Available: airc.it/cancro/informazioni-tumori/guida-ai-tumori/tumore-del-seno.

[6] A. Raghavendra, A. K. Sinha and H. T. Le-Petross et al., "Mammographic breast density is associated with the development of contralateral breast cancer," *Cancer,* vol. 123 (11), pp. 1935-1940, 2017.

[7] Memorial Sloan Kettering Cancer Center, "Anatomy of the Breast," 2019. [Online]. Available: mskcc.org/cancer-care/types/breast/anatomy-breast.

[8] Breastcancer.org, «Types of Breast Cancer,» 2018. [Online]. Available: breastcancer.org/symptoms/types.

[9] National Breast Cancer Foundation (NBCF), "Learn about Breast Cancer," 2019. [Online]. Available: nationalbreastcancer.org/about-breast-cancer/.

[10] A. Casasent, M. Edgerton and N. E. Navin, "Genome evolution in ductal carcinoma in situ: invasion of the clones," *The Journal of Pathology,* vol. 241 (2), pp. 208-218, 2017.

[11] American College of Radiology (ACR), «ACR BI-RADS Atlas 5th Edition,» 2013. [Online]. Available: acr.org/-/media/ACR/Files/RADS/BI-RADS/BIRADS-Reference-Card.pdf.

[12] R. M. Rangayyan, N. R. Mudigonda and J. E. L. Desautels, "Boundary modelling and shape analysis methods for classification of mammographic masses," *Medical & Biological Engineering & Computing,* vol. 38, 2000.

[13] Mayo Foundation for Medical Education and Research (MFMER), "Breast Cancer," Mayo Clinic, 2019. [Online]. Available: mayoclinic.org/diseases-conditions/breast-cancer/diagnosis-treatment/drc-20352475.

[14] S. Kul, S. Oguz, I. Eyuboglu e O. Komurcuoglu, «Can unenhanced breast MRI be used to decrease negative biopsies rates?,» *Diagnostic and Interventional Radiology,* vol. 21 (4), pp. 287-292, 2015.

[15] J. S. Drukteinis, B. P. Mooney, C. I. Flowers and R. A. Gatenby, "Beyond mammography: new frontiers in breast cancer screening," *The American Journal of Medicine,* vol. 126(6), pp. 472-479, 2013.

[16] V. S. Subbhuraam, E. Ng Yin Kwee, U. A. Rajendra and F. Oliver, "Breast imaging: A survey," *World Journal of Clinical Oncology,* vol. 2 (4), pp. 171-178, 2011.

[17] M. Hammer, "X-Ray Physics: X-Ray Interaction with Matter and Attenuation," 2014. [Online]. Available: xrayphysics.com/attenuation.html.

[18] J. M. Seely and T. Alhassan, "Screening for breast cancer in 2018 - what should we be doing today?," *Current Oncology,* vol. 25 (S1), pp. 115-124, 2018.

[19] K. Holland, I. Sechopoulos, R. M. Mann, G. J. den Heeten, C. H. van Gils and N. Karssemeijer, "Influence of breast compression pressure on the performance of population-based mammography screening," *Breast Cancer Research,* vol. 19, 2017.

[20] National Cancer Institute, "PDQ Cancer Information Summaries," 2016. [Online]. Available: ncbi.nlm.nih.gov/books/NBK65715.3.

[21] E. Samei, J. Thompson, S. Richard and J. Bowsher, "A Case for Wide-Angle Breast Tomosynthesis," *Academic Radiology,* vol. 22 (7), pp. 860-869, 2015.

[22] S. S. J. Feng and I. Sechopoulos, "Clinical Digital Breast Tomosynthesis System: Dosimetric Characterization," *Radiology,* vol. 263 (1), pp. 35-42, 2012.

[23] E. F. Conant, "Clinical Implementation of Digital Breast Tomosynthesis," *Radiologic Clinics of North America,* vol. 52 (3), pp. 499-518, 2014.

[24] T. Nguyen, G. Levy and E. Poncelet et al., "Overview of digital breast tomosynthesis: Clinical cases, benefits and disadvantages," *Diagnostic and Interventional Imaging,* vol. 96 (9), pp. 843-859, 2015.

[25] E. Devolli-Disha, S. Manxhuka-Kerliu, H. Ymeri and A. Kutllovci, "Comparative Accuracy of Mammography and Ultrasound in Women with Breast Symptoms according to Age and Breast Density," *Bosnian Journal of Basic Medical Sciences,* vol. 9 (2), pp. 131-136, 2009.

[26] R. F. Brem, M. J. Lenihan, J. Lieberman e J. Torrente, «Screening Breast Ultrasound: Past, Present, and Future,» *American Journal of Roentgenology,* vol. 204, pp. 234-240, 2015.

[27] D. Thigpen, A. Kappler e R. Brem, «The Role of Ultrasound in Screening Dense Breasts - A Review of the Literature and Practical Solutions for Implementation,» *DIagnostics (Basel),* vol. 8 (1), 2018.

[28] F. Taskin, Y. Polat and I. H. Erdogdu et al., "Problem-solving breast MRI: useful or a source of new problems," *Diagnostic and Interventional Radiology,* vol. 24 (5), pp. 255-261, 2018.

[29] G. L. G. Menezes, F. M. Knuttel and B. L. Stehouwer et al., "Magnetic resonance imaging in breast cancer: A literature review and future perspectives," *World Journal of Clinical Oncology,* vol. 5 (2), pp. 61-70, 2014.

[30] Johns Hopkins Medicine, «Breast Magnetic Resonance Imaging (MRI),» 2019. [Online]. Available: www.hopkinsmedicine.org/health/treatment-tests-and-therapies/breast-mri.

[31] J. G. Elmore, K. Armstrong, C. D. Lehman and S. W. Fletcher, "Screening for Breast Cancer," *Journal of the American Medical Association (JAMA),* vol. 293 (10), pp. 1245-1256, 2005.

[32] American Cancer Society, "Breast MRI," 2017. [Online]. Available: www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/breast-mri-scans.html.

[33] Koning Corporation, "Koning Revolutionary Technology," 2019. [Online]. Available: koninghealth.com/en/technology/.

[34] K. K. Lindfors, J. M. Boone, M. S. Newell and C. J. D'Orsi, "Dedicated Breast CT: The Optimal Cross Sectional Imaging Solution?," *Radiologic clinics of North America,* vol. 48 (5), pp. 1043-1054, 2010.

[35] K. K. Lindfors, J. M. Boone and T. R. Nelson et al., "Dedicated Breast CT: Initial Clinical Experience," *Radiology,* vol. 246 (3), pp. 725-733, 2008.

[36] H. C. Kuo, M. L. Giger, I. Reiser, J. M. Boone and K. K. Lindfors et al., "Level Set Segmentation of Breast Masses in Contrast-Enhanced Dedicated Breast CT and Evaluation of Stopping Criteria," *Journal of Digital Imaging,* vol. 27 (2), pp. 237-247, 2014.

[37] M. Caballo, J. M. Boone, R. Mann and I. Sechopoulos, "An unsupervised automatic segmentation algorithm for breast tissue classification of dedicated breast computed tomography images," *Medical Physics,* vol. 45 (6), pp. 2542-2559, 2018.

[38] H. K. Jung, C. M. Kuzmiak, K. W. Kim and N. M. Choi et al., "Potential Use of American College of Radiology BI-RADS Mammography Atlas for Reporting and Assessing Lesions Detected on Dedicated Breast CT

Imaging: Preliminary Study," *Academic Radiology,* vol. 24 (11), pp. 1395-1401, 2017.

[39] M. Caballo, J. Teuwen, R. M. Mann and I. Sechopoulos, "Breast parenchyma analysis and classification for breast masses detection using texture feature descriptors and neural networks in dedicated breast CT images," *Proceeding SPIE Medical Imaging 2019: Computer-Aided Diagnosis,* vol. 10950, 2019.

[40] N. Kiarashi and E. Samei, "Digital breast tomosynthesis: a concise overview," *Imaging in Medicine,* vol. 5 (5), 2013.

[41] E. Garcia, Y. Diez and X. Llado et al., "Breast MRI and X-ray mammography registration using gradient values," *Medical Image Analysis,* vol. 54, pp. 76-87, 2019.

[42] J. M. Boone, T. R. Nelson, K. K. Lindfors and J. A. Seibert, "Dedicated breast CT: radiation dose and image quality evaluation," *Radiology,* vol. 221 (3), pp. 657-667, 2001.

[43] A. M. O'Connell, A. Karellas e S. Vedantham, «The potential role of dedicated 3D breast CT as a diagnostic tool: Review and early clinical examples,» *The Breast Journal,* vol. 20 (6), pp. 592-605, 2014.

[44] I. Sechopoulos, S. S. J. Feng and C. J. D'Orsi, "Dosimetric characterization of a dedicated breast computed tomography clinical prototype," *Medical Physics,* vol. 37 (8), pp. 4110-4120, 2010.

[45] M. Caballo, C. Fedon, L. Brombal, R. Mann, R. Longo and I. Sechopoulos, "Development of 3D patient-based super-resolution digital breast phantoms using machine learning," *Physics in Medicine and Biology,* vol. 63 (22), 2018.

[46] R. J. Gilles, P. E. Kinahan e H. Hricak, «Radiomics: Images Are More than Pictures, They Are Data,» *Radiology,* vol. 278 (2), pp. 563-577, 2013.

[47] P. Afshar and A. Mohammadi et al., "From Hand-Crafted to Deep Learning-based Cancer Radiomics: Challenges and Opportunities," *IEEE Signal Processing Magazine,* vol. 36 (4), pp. 132 - 160, 2019.

[48] P. Lambin, E. Rios-Velazquez and R. Leijenaar et al., "Radiomics: Extracting more information from medical images using advanced feature analysis," *European Journal of Cancer,* vol. 48 (4), pp. 441 - 446, 2012.

[49] Advanced X-ray Tomographic Imaging (AXTI) Laboratory, "Automated breast cancer detection and characterization in dedicated breast CT imaging," 2019. [Online]. Available: axti.radboudimaging.nl/index.php/Dedicated_breast_CT.

[50] V. Kumar, Y. Gu, S. Basu and A. Berglund et al., "QIN "Radiomics: The Process and the Challenges"," *Magnetic Resonance Imaging,* vol. 30 (9), pp. 1234 - 1248, 2012.

[51] R. M. Haralick, K. Shanmugam and I. H. Dinstein, "Textural Features for Image Classification," *IEEE Transactions on Systems, Man, and Cybernetics,* Vols. SMC-3 (6), pp. 610 - 621, 1973.

[52] M. M. Galloway, "Texture analysis using gray level run lengths," *Computer Graphics and Image Processing,* vol. 4 (2), pp. 172 - 179, 1975.

[53] T. Ojala, M. Pietikainen and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 24 (7), pp. 971 - 987, 2002.

[54] D. Zhang, A. Wong, M. Indrawan and G. Lu, "Content-based Image Retrieval Using Gabor Texture Features," *IEEE Transactions PAMI,* pp. 13 - 15, 2000.

[55] A. F. Pinheiro Sequeira, "Liveness Detection and Robust Recognition in Iris and Fingerprint Biometric Systems," Faculdade de Engenharia da Universidade do Porto, Porto, 2015.

[56] K. Preetha and S. K. Jayanthi, "GLCM and GLRLM based Feature Extraction Technique in Mammogram Images," *International Journal of Engineering and Technology,* vol. 7 (2.21), pp. 266 - 270, 2018.

[57] J. Liu and Y. Shi, "Image Feature Extraction Method Based on Shape Characteristics and Its Application in Medical Image Analysis," *International Conference on Applied Informatics and Communication (ICAIC),* vol. CCIS 224, pp. 172 - 178, 2011.

[58] C. Shen, Z. Liu, M. Guan and J. Song et al., "2D and 3D CT Radiomics Features Prognostic Performance Comparison in Non-Small Cell Lung Cancer," *Translational Oncology,* vol. 10 (6), pp. 886 - 894, 2017.

[59] Y. Zhang, A. Oikonomou and A. Wong et al., "Radiomics-based Prognosis Analysis for Non-Small Cell Lung Cancer," *Scientific Reports,* vol. 7, 2017.

[60] H. J. W. L. Aerts, E. R. Velazquez and R. T. H. Leijenaar et al., "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature Communications,* vol. 5, 2014.

[61] J. J. M. Griethuysen, A. Fedorov and C. Parmar et al., "Computational Radiomics System to Decode the Radiographic Phenotype," *Cancer Research,* vol. 77 (21), pp. 104 - 107, 2017.

[62] A. Vignati, S. Mazzetti, V. Giannini and F. Russo et al., "Texture features on T2-weighted magnetic resonance imaging: new potential biomarkers for prostate cancer aggressiveness.," *Physics in Medicine and Biology,* vol. 60 (7), pp. 2685 - 2701, 2015.

[63] A. Wibmer, H. Hricak, T. Gondo and K. Matsumoto et al., "Haralick texture analysis of prostate MRI: utility for differentiating non-cancerous prostate from prostate cancer and differentiating prostate cancers with

different Gleason scores.," *European Radiology,* vol. 25 (10), pp. 2840 - 2850, 2015.

[64] Y. Zhou, L. He e Y. Huang et al., «CT-based radiomics signature: a potential biomarker for preoperative prediction of early recurrence in hepatocellular carcinoma.,» *Abdominal Radiology (New York),* vol. 42 (6), pp. 1695 - 1704, 2017.

[65] N. Mao, P. Yin, Q. Wang and M. Liu et al., "Added Value of Radiomics on Mammography for Breast Cancer Diagnosis: A Feasibility Study," *Journal of the American College of Radiology,* vol. 16 (4), pp. 485 - 491, 2019.

[66] J. Li, Y. Song, S. Xu and J. Wang et al., "Predicting underestimation of ductal carcinoma in situ: a comparison between radiomics and conventional approaches," *International Journal of Computer Assisted Radiology and Surgery,* vol. 14 (4), pp. 709 - 721, 2019.

[67] H. P. Chan, Y. T. Wu and B. Sahiner et al., "Characterization of masses in digital breast tomosynthesis: Comparison of machine learning in projection views and reconstructed slices," *Medical Physics,* vol. 37 (7), pp. 3576 - 3586, 2010.

[68] S. E. Lee, K. Han and J. Y. Kwak et al., "Radiomics of US texture features in differential diagnosis between triple-negative breast cancer and fibroadenoma," *Scientific Reports,* vol. 8, 2018.

[69] H. A. Nugroho, Y. Triyani, M. Rahmawaty e I. Ardiyanto, «Analysis of Margin Sharpness for Breast Nodule Classification on Ultrasound Images,» *9th International Conference on Information Technology and Electrical Engineering (ICITEE),* 2017.

[70] S. Huang, B. L. Franc, R. J. Harnish and G. Liu et al., "Exploration of PET and MRI radiomic features for decoding breast cancer phenotypes and prognosis," *NPJ Breast Cancer,* vol. 4 (24), 2018.

[71] Q. Xiong, X. Zhou, Z. Liu e C. Lei et al., «Multiparametric MRI-based radiomics analysis for prediction of breast cancers insensitive to neoadjuvant chemotherapy.,» *Clinical & Translational Oncology,* 2019.

[72] I. Reiser, R. M. Nishikawa, M. L. Giger, J. M. Boone, K. K. Lindfors and K. Yang, "Automated detection of mass lesions in dedicated breast CT: a preliminary study," *Medical Physics,* vol. 39 (2), pp. 866 - 873, 2012.

[73] S. Ray, N. D. Prionas, K. K. Lindfors e J. M. Boone, «Analysis of breast CT lesions using computer-aided diagnosis: an application of neural networks on extracted morphologic and texture features,» *Proceeding SPIE Medical Imaging 2012: Computer-Aided Diagnosis,* vol. 8315, 2012.

[74] Centro Matematica - Unità Città Studi (Dipartimento di Matematica), "Piani di simmetria di un cubo," Immagini per la Matematica, 2019. [Online]. Available: matematita.it/materiale/?p=cat&sc=271,272,387&im=1098.

[75] B. Surendiran and A. Vadivel, "Mammogram mass classification using various geometric shape and margin features for early detection of breast cancer," *International Journal of Medical Engineering and Informatics,* vol. 4 (1), pp. 36 - 54, 2012.

[76] C. Cheng, W. Liu e H. Zhang, «Image retrieval based on region shape similarity,» *Proceedings SPIE, Storage and Retrieval for Media Databases,* vol. 4315, pp. 31 - 37, 2001.

[77] M. Peura and J. Iivarinen, "Efficiency of simple shape descriptors," *Proceedings of the 3rd International Workshop on Visual Form (IWVF),* pp. 443 - 451, 1997.

[78] J. Kilday, F. Palmieri and M. D. Fox, "Classifying Mammographic Lesions using Computerized Image Analysis," *IEEE Transactions on Medical Imaging,* vol. 12 (4), pp. 664 - 669, 1993.

[79] L. Shen, R. M. Rangayyan and J. E. L. Desautels, "Application of Shape Analysis to Mammographic Calcifications," *IEEE Transactions on Medical Imaging,* vol. 13 (2), pp. 263 - 274, 1994.

[80] L. Gupta and M. D. Srinath, "Contour sequence moments for the classification of closed planar shapes," *Pattern Recognition,* vol. 20 (3), pp. 267 - 272, 1987.

[81] K. Kpalma and J. Ronsin, "Multiscale contour description for pattern recognition," *Pattern Recognition Letters,* vol. 27 (13), pp. 1545 - 1559, 2006.

[82] Z. Huo, M. L. Giger, Vyborny et al. and C. J. Vyborny et al., "Analysis of spiculation in the computerized classification of mammographic masses," *Medical Physics,,* vol. 22 (10), pp. 1569 - 1579, 1995.

[83] Z. Huo, M. L. Giger and C. J. Vyborny et al., "Automated Computerized Classification of Malignant and Benign Masses on Digitized Mammograms," *Academic Radiology,* vol. 5 (3), pp. 155 - 168, 1998.

[84] C. M. Sehgal, T. W. Cary and S. A. Kangas et al., "Computer-Based Margin Analysis of Breast Sonography for Differentiating Malignant and Benign Masses," *Journal of Ultrasound in Medicine,* vol. 23, pp. 1201 - 1209, 2004.

[85] J. Xu, S. Napel and H. Greenspan et al., "Quantifying the margin sharpness of lesions on radiological images for content-based image retrieval," *Medical Physics,* vol. 39 (9), pp. 5405 - 5418, 2012.

[86] E. Alpaydin, Introduction to Machine Learning, Cambridge (Massachusetts): The MIT Press, 2010.

[87] Four Years Remaining, "The Mystery of Early Stopping," 2017. [Online]. Available: fouryears.eu/2017/12/06/the-mystery-of-early-stopping/.

[88] J. Lao, Y. Chen and Z. Li et al., "A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme," *Scientific Reports,* vol. 7 (10353), 2017.

[89] M. A. Hall, "Correlation-based Feature Selection for Machine Learning," Department of Computer Science at The University of Waikato, 1999. [Online]. Available: cs.waikato.ac.nz/~mhall/thesis.pdf.

[90] N. P. Pérez, M. A. G. Lopez, A. Silva and I. Ramos, "Improving the Mann–Whitney statistical test for feature selection: An approach in breast cancer diagnosis on mammography," *Artificial Intelligence in Medicine,* vol. 63, pp. 19 - 31, 2015.

[91] L. Breiman, J. Friedman, C. J. Stone and R. A. Olshen, Classification and Regression Trees, Chapman and Hall/CRC, 1984.

[92] W. Y. Loh and Y. S. Shih, "Split Selection Methods for Classification Trees," *Statistica Sinica,* vol. 7, pp. 815-840, 1997.

[93] S. Kotsiantis, D. Kanellopoulos and P. Pintelas, "Handling Imbalanced Datasets: A Review," *GESTS International Transactions on Computer Science and Engineering,* vol. 30, 2006.

[94] M. Sciences and N. Winovich, "Neural Networks," Mathematical Sciences - Purdue University, 2019. [Online]. Available: math.purdue.edu/~nwinovic/deep_learning.html.

[95] M. F. Moller, "A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning," *Neural Networks,* vol. 6, pp. 525-533, 1993.

[96] J. Davis e M. Goadrich, «The relationship between Precision-Recall and ROC curves,» *ICML '06 Proceedings of the 23rd International Conference on Machine Learning,* pp. 233-240, 2006.