UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE FÍSICA

Breast Cancer: Automatic Detection and Risk Analysis through Machine Learning Algorithms, using mammograms

João Pedro Velhinho Mendes

Mestrado Integrado em Engenharia Biomédica e Biofísica
Engenharia Clínica e Instrumentação Médica

Dissertação orientada por:
Professor Doutor Nuno Matela

2021

Acknowledgments

Em primeiro lugar, gostaria de agradecer ao meu orientador, Professor Nuno Matela: pelo apoio desde o primeiro dia em que mostrei interesse pelo tema que culminou neste trabalho, pelas respostas a todas as dúvidas e inseguranças que tive, por todas as contribuições valiosas que são parte integrante deste documento, e por tantas vezes me fazer ver que os problemas que tinha eram, invariavelmente, insignificantes.

Depois, queria agradecer a todos aqueles que, de uma forma ou de outra, se cruzaram comigo ao longo do decorrer deste trabalho e que me deixaram o seu apoio. De todos, destaco quatro: Quero agradecer ao Duarte, por tantas vezes se ter mostrado tão entusiasmado como eu pelo projeto, motivando-me ainda mais. Quero agradecer ao Vasco, pelas anotações que foi deixando, pelas sugestões que foi fazendo, e pelos abraços que foi dando quando as coisas corriam menos bem. Quero agradecer ao Manuel, pelo apoio que me dá a cada pequena vitória mesmo não fazendo ideia do que se está a passar. E, claro, quero agradecer à Sara por tantas coisas que um parágrafo não chega para as conter.. nomeio algumas: pelo interesse pelo meu trabalho, pelas revisões e anotações milimétricas, pela paciência que tem para mim, e por ser a aberta nas tempestades em copos de água que tantas vezes fiz.

Quero agradecer ao meu Tio Fernando, à minha Tia Angélica, e à minha Avó Joaquina, pelo apoio fundamental que me deram ao longo deste ano, e que tornou a minha vida muito mais tranquila.

Por fim, mas não menos importante, quero agradecer à minha família mais próxima: aos meus pais por, no meio de tantas dificuldades, nunca terem abdicado de investir na minha educação, por nunca terem exigido de mim mais do que dar o meu melhor, e por toda a compreensão e apoio que me deram em todas as decisões, mesmo que não fossem as melhores; ao meu irmão que, mesmo sem o verbalizar, me apoia todos os dias, sabendo que as vitórias, minhas e dele, serão sempre partilhadas… e por gozar comigo a toda a hora e me fazer perceber que não sou assim tão importante.

Quero, em última instância, agradecer à minha Avó Rosa. O amor, a amizade, a luta pelo que acreditamos, a resiliência, o rigor, e o "brio", são valores que ela me passava em todas as conversas que tínhamos, mesmo naquelas em que só ela é que falava. E a avó Rosa, mesmo não podendo assistir ao cumprir de mais um objetivo, está, de uma forma ou de outra, patente ao longo deste trabalho.

Abstract

Two million and three hundred thousand Breast Cancer (BC) cases were diagnosed in 2020, making it the type of cancer with the highest incidence that year, considering both sexes. Breast Cancer diagnosis usually occurs during screening programs using mammography, which has some downsides: the masking effect due to its 2-D nature, and its poor sensitivity concerning dense breasts. Since these issues result in difficulties reading mammograms, the main part of this work aimed to verify how a computer vision method would perform in classifying mammograms into two classes: cancer and non-cancer. The 'non-cancer group' (N=159) was composed by images with healthy tissue (N=84) and images with benign lesions (N=75), while the cancer group (N=73) contained malignant lesions. To achieve this, multiple classifiers were optimized and trained ($N_{train} = 162, N_{test} = 70$) with a previously selected ideal sub-set of features that describe the texture of the entire image, instead of just one small Region of Interest (ROI). The classifier with the best performance was Support Vector Machine (SVM), (AUC = 0.875), which indicates a good-to-excellent capability discriminating the two defined groups. To assess if Percent Mammographic Density (%PD), an important risk factor, added important information, a new classifier was optimized and trained using the selected sub-set of texture features plus the %PD calculation. The classifier with the best performance was a Linear Discriminant Analysis (LDA), (AUC=0.875), which seems to indicate, once it achieves the same performance as the classifier using only texture features, that there is no relevant information added from %PD calculations. This happens because texture already includes information on breast density. To understand how the classifier would perform in worst image acquisition conditions, gaussian noise was added to the test images (N=70), with four different magnitudes (AUC= 0.765 for the lowest noise value vs. AUC ≈ 0.5 for the highest). A median filter was applied to the noised images towards evaluating if information could be recovered. For the highest noise value, after filtering, the AUC was very close to the one obtained for the lowest noise value before filtering (0.754 vs 0.765), which indicates information recovery. The effect of density in classifier performance was evaluated by constructing three different test sets, each containing images from a density class (1,2,3). It was seen that an increase in density did not necessarily resulted in a decrease in performance, which indicates that the classifier is robust to density variation (AUC = 0.864, AUC= 0.927, AUC= 0.905 ; for class 1, 2, and 3 respectively). Since the entire image is being analyzed, and images come from different datasets, it was verified if breast area was adding bias to classification. Pearson correlation coefficient provided an output of $\rho = 0.22$, showing that there is a weak correlation between these two variables. Finally, breast cancer risk was assessed by visual texture feature analysis through the years, for a small set of women (N=11). This visual analysis allowed to unveil what seems to be a pattern amongst women who developed the disease, in the mammogram immediately before diagnosis. The details of each phase, as well as the associated final results are deeply described throughout this document. The work done in the first classification task resulted in a state-of-the-art performance, which may serve as foundation for new research in the area, without the laborious work of ROI definition. Besides that, the use of texture features alone proved to be fruitful. Results concerning risk may serve as basis for future work in the area, with larger datasets and the incorporation of Computer Vision methods.

Keywords: Breast Cancer, Medical Imaging, Texture Features, Machine Learning, Risk Analysis

Resumo

Com 2.3 milhões de casos diagnosticados em todo o Mundo, durante o ano de 2020, o cancro da mama tornou-se aquele com maior incidência, nesse mesmo ano, considerando ambos os sexos. Anualmente, em Portugal, são diagnosticados aproximadamente sete mil (7000) novos casos de cancro da mama, com mil oitocentas (1800) mulheres a morrerem, todos os anos, devido a esta doença - indicando uma taxa de mortalidade de aproximadamente 5 mulheres por dia. A maior parte dos diagnósticos de cancro da mama ocorrem ao nível de programas de rastreio, que utilizam mamografia. Esta técnica de imagem apresenta alguns problemas: o facto de ser uma imagem a duas dimensões leva a que haja sobreposição de tecidos, o que pode mascarar a presença de tumores; e a fraca sensibilidade a mamas mais densas, sendo estas caraterísticas de mulheres com risco de cancro da mama mais elevado. Como estes dois problemas dificultam a leitura das mamografias, grande parte deste trabalhou focou-se na verificação do desempenho de métodos computacionais na tarefa de classificar mamografias em duas classes: cancro e não-cancro. No que diz respeito à classe "não-cancro" (N = 159), esta foi constituída por mamografias saudáveis (N=84), e por mamografias que continham lesões benignas (N=75). Já a classe "cancro" continha apenas mamografias com lesões malignas (N = 73). A discriminação entre estas duas classes foi feita com recurso a algoritmos de aprendizagem automática. Múltiplos classificadores foram otimizados e treinados ($N_{treino} = 162, N_{teste} = 70$), recorrendo a um conjunto de características previamente selecionado, que descreve a textura de toda a mamografia, em vez de apenas uma única Região de Interesse. Estas características de textura baseiam-se na procura de padrões: sequências de pixéis com a mesma intensidade, ou pares específicos de pixéis. O classificador que apresentou uma performance mais elevada foi um dos *Support Vector Machine* (SVM) treinados – AUC= 0.875, o que indica um desempenho entre o bom e o excelente. A *Percent Mammographic Density* (%PD) é um importante fator de risco no que diz respeito ao desenvolvimento da doença, pelo que foi estudado se a sua adição ao set de features selecionado resultaria numa melhor performance dos classificadores. O classificador, treinado e otimizado utilizando as features de textura e os cálculos de %PD, com maior capacidade discriminativa foi um *Linear Discriminant Analysis* (LDA) – AUC = 0.875. Uma vez que a performance é igual à obtida com o classificador que utiliza apenas features de textura, conclui-se que a %PD parece não contribuir com informação relevante. Tal pode ocorrer porque as próprias características de textura já têm informação sobre a densidade da mama. De forma a estudar-se de que modo o desempenho destes métodos computacionais pode ser afetado por piores condições de aquisição de imagem, foi simulado ruído gaussiano, e adicionado ao set de imagens utilizado para testagem. Este ruído, adicionado a cada imagem com quatro magnitudes diferentes, resultou numa AUC de 0.765 para o valor mais baixo de ruído, e numa AUC de 0.5 para o valor de ruído mais elevado. Tais resultados indicam que, para níveis de ruído mais baixo, o classificador consegue, ainda assim, manter uma performance satisfatória – o que deixa de se verificar para valores mais elevados de ruído. Estudou-se, também, se a aplicação de técnicas de filtragem – com um filtro mediana – poderia ajudar a recuperar informação perdida aquando da adição de ruído. A aplicação do filtro a todas as imagens ruidosas resultou numa AUC de 0.754 para o valor mais elevado de ruído, atingindo assim um desempenho similar ao set de imagens menos ruidosas, antes do processo de filtragem (AUC=0.765). Este resultados parecem indicar que, na presença de más condições de aquisição, a aplicação de um filtro mediana pode ajudar a recuperar informação, conduzindo assim a um melhor desempenho dos métodos computacionais. No entanto, esta mesma conclusão parece não se verificar para valores de ruído mais baixo onde a AUC após filtragem acaba por ser mais reduzida. Tal resultado poderá indicar que, em situações onde o nível de ruído é mais baixo, a técnica de filtragem não só remove o ruído, como acaba também por, ela própria, remover informação ao nível da textura

da imagem. De modo a verificar se mamas com diferentes densidades afetavam a performance do classificador, foram criados três sets de teste diferentes, cada um deles contendo imagens de mamas com a mesma densidade (1, 2, e 3). Os resultados obtidos indicam-nos que um aumento na densidade das mamas analisadas não resulta, necessariamente, numa diminuição da capacidade em discriminar as classes definidas (AUC = 0.864, AUC = 0.927, AUC= 0.905; para as classes 1, 2, e 3 respetivamente). A utilização da imagem integral para analisar de textura, e a utilização de imagens de datasets diferentes (com dimensões de imagem diferentes), poderiam introduzir um viés na classificação, especialmente no que diz respeito às diferentes áreas da mama. Para verificar isso mesmo, utilizando o coeficiente de correlação de Pearson, $\rho = 0.3$, verificou-se que a área da mama (e a percentagem de ocupação) tem uma fraca correlação com a classificação dada a cada imagem. A construção do classificador, para além de servir de base a todos os testes apresentados, serviu também o propósito de criar uma interface interativa, passível de ser utilizada como ficheiro executável, sem necessidade de instalação de nenhum software. Esta aplicação permite que o utilizador carregue imagens de mamografia, exclua *background* desnecessário para a análise da imagem, extraia features, teste o classificador construído e dê como output, no ecrã, a classe correspondente à imagem carregada. A análise de risco de desenvolvimento da doença foi conseguida através da análise visual da variação dos valores das features de textura ao longo dos anos para um pequeno set (N=11) de mulheres. Esta mesma análise permitiu descortinar aquilo que parece ser uma tendência apresentada apenas por mulheres doentes, na mamografia imediatamente anterior ao diagnóstico da doença. Todos os resultados obtidos são descritos profundamente ao longo deste documento, onde se faz, também, uma referência pormenorizada a todos os métodos utilizados para os obter. O resultado da classificação feita apenas com as features de textura encontra-se dentro dos valores referenciados no estado-da-arte, indicando que o uso de features de textura, por si só, demonstrou ser profícuo. Para além disso, tal resultado serve também de indicação que o recurso a toda a imagem de mamografia, sem o trabalho árduo de definição de uma Região de Interesse, poderá ser utilizado com relativa segurança. Os resultados provenientes da análise do efeito da densidade e da área da mama, dão também confiança no uso do classificador. A interface interativa que resultou desta primeira fase de trabalho tem, potencialmente, um diferenciado conjunto de aplicações: no campo médico, poderá servir de auxiliar de diagnóstico ao médico; já no campo da análise computacional, poderá servir para a definição da *ground truth* de potenciais datasets que não tenham legendas definidas. No que diz respeito à análise de risco, a utilização de um dataset de dimensões reduzidas permitiu, ainda assim, compreender que existem tendências nas variações das features ao longo dos anos, que são especificas de mulheres que desenvolveram a doença. Os resultados obtidos servem, então, de indicação que a continuação desta linha de trabalho, procurando avaliar/predizer o risco, deverá ser seguida, com recurso não só a datasets mais completos, como também a métodos computacionais de aprendizagem automática.

Palavras-Chave: Cancro da Mama, Imagem Médica, Características de Textura, Aprendizagem Automática, Análise de Risco

Table of Contents

List of Figures

List of Tables

List of Abbreviations

%PD	Percent Mammographic Density
ACR	American College of Radiology
AUC	Area Under the Curve
BC	Breast Cancer
BI-RADS	Breast Imaging-Reporting and Data System
BMI	Body Mass Index
CBIS-DDSM	Curated Breast Imaging Subset of Digital Database for Screening Mammography
CC	Cranio-Caudal
DA	Discriminant Analysis
FDA	Food and Drug Administration
FMP	First Moment of the Power Spectrum
FPR	False Positive Rate
GLCM	Gray Level Co-Occurrence Matrix
k-NN	k Nearest Neighbor
LBP	Local Binary Pattern
LCIS	Lobular Carcinoma *in situ*
LR	Logistic Regression
MI	Mutual Information
ML	Machine Learning
MLO	Medio-lateral Oblique
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
NGTDM	Neighbourhood Gray-Tone Difference Matrix
pSNR	Peak Signal-to-Noise Ratio
RLM	Run-Length Matrix
RMS	Root Means Square of the Power Spectrum
ROC	Relative Operating Characteristic
ROI	Region of Interest
SS	Structural Similarity
SVM	Support Vector Machine
TPR	True Positive Rate
US	Ultrasound

1. Introduction

Section 1 is divided into nine subsections. The first gives a motivation for the developed work, showing worldwide data concerning Breast Cancer (BC). Section 1.2 offers a clinical background, relate to breast anatomy, which is complemented in section 1.3, where breast diseases are analyzed. In subsection 1.4, genetic and environmental risk factors for the development of this disease are described, while in subsection 1.5 screening methodologies, and their advantages and disadvantages, are examined. The information studied in the previous referred subsection is deepened in subsection 1.6, where mammography is profoundly explored, in term of its physical principles, as in terms of the most common findings in this type of exam. Subsections 1.7 and 1.8 are more related to the purpose of these work, with feature extraction and machine learning (ML) introductions being made, respectively. Finally, subsection 1.9 presents the primary goals of this thesis.

1.1 – Motivation

One in eight women will be diagnosed with BC in their lifetime, with one in thirty-nine women dying from this disease, only in the United States of America. In the same country, in 2020, approximately 42 170 women were expected to die from BC and it was anticipated that approximately 30% of the cancers detected in women were BC [1]. Around 95% of cancers are due to genetic mutations that result from environmental or lifestyle factors, where the remaining percentage is related to inherited genes – with BRCA1/BRCA2 genes being responsible for most of cases of BC [2, 3].

BC diagnosis occurs either during a common screening program, before symptoms appear, or after women noticing breast changes. Screening programs are important for an early detection of BC - that is, in a more treatable stage - resulting in a decrease in mortality [1, 4].

The criterion that defines if a woman is eligible for screening is, normally, her age. Different countries have different recommendations on which age is the best to start screening; the USA states that women from age 45 to 54 should have a mammography once a year, while 55+ plus women should have a mammography once every two years. On the other hand, the UK National Health System says that only women between 50 to 71 should be screened, and only once every three years [5, 6].

Incidence rates of BC increased highly during the decades of 1980 and 1990, mostly due to the increase in mammography screening programs available – studies even point out a prevalence increase from 29% to 70% resulting from screening within the time frame from 1987 to 2000 [7]. This increase can be seen, for the case of invasive BC, in Figure 1.1. The recent increase in BC incidence is thought to be related to a higher Body Mass Index (BMI) and a diminished number of births per woman [8].

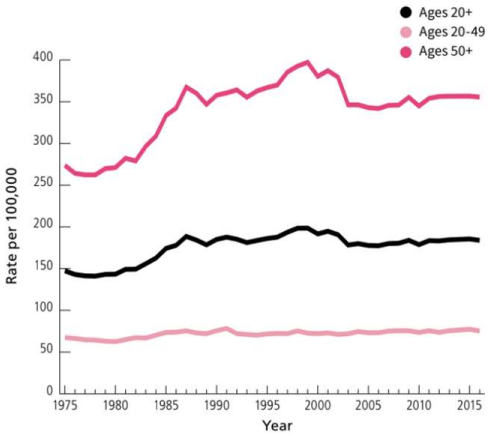

Figure 1.1 - Incidence Rate of Invasive Breast Cancer in the United States of America. [1]

In what concerns to the mortality of this disease, an extensive decline has been observed, with a drop of nearly 40% from 1989 to 2017, although having slowed down in recent years. This decline in mortality may be related not only to better treatments but also with best preventive tools [9]. Nonetheless, not all women were affected by this improvement in both diagnostic and treatment options, as it can be seen in Figure 1.2, with the descending trend in Non-Hispanic Black Woman not being has leaned as it happens in Non-Hispanic White Women. There are a great number of factors contributing for this difference, and while one might argue that there are unfortunate tumor characteristics specific to this ethnicity, social disadvantages must not be overlooked. Black women tend to have less access to high-quality diagnostic and treatment options, having a higher probability of being screened in institutions that don't have as many resources or are even nonaccredited at all. This can lead not only to a long period of time between examination and getting the results, as also to a poor assessment and consequently to a bad follow-up [1].

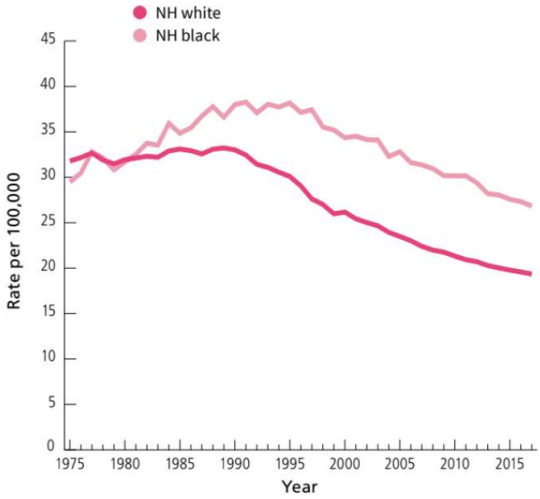

Figure 1.2 - Mortality Rates of Breast Cancer between Non-Hispanic white Women and Non-Hispanic Black Women. [1]

The treatment for invasive BC involves a panoply of options that drive from surgery, like breast removal (mastectomy), to radiation therapy, that is normally done after surgery to eliminate abnormal cell that remained in the breast area. Besides that, chemotherapy; hormonal therapy, to block the effects of estrogen, that stimulates BC; and Immunotherapy, that is the use of drugs to stimulate one's immune system to recognize and destroy cancer cells, are other possible lines of treatment [1].

Although there are multiple screening programs, they might not serve all women. Some younger women may be at higher risk of developing BC than women in their fifties and, despite that, these women are not eligible for screening. With that in mind, the perfect screening program should not consider age as the *only* risk factor that determines when to screen women.

The current medicine paradigm is one of preventive and personalized care and, for that reason, methodologies that allow the prevention, or an early diagnosis, of this disease, in a personalized fashion, are highly valuable. Although screening programs aim to prevent further effects of developing BC by making an early diagnosis, different factors like screening periodicity, may disrupt this goal. A technique that allowed an automatic detection of the disease, using a mammogram, would aid the doctors in the diagnostic process, making it easier and faster. An application like this one would be of great importance, since time is imperative in what concerns to cancerous diseases diagnosis and treatment. Besides that, a technology that allowed, based on a mammogram and other epidemiological factors, to retrieve a risk of future development of BC, would enable doctors to, in a personalized fashion, determine how and when to screen women, and if some preventive course of treatment should be made.

1.2 – Breast Anatomy

The breasts are an accumulation of tissue that is located in the pectoral region, on the anterior thoracic wall, overlying the *pectorales major* – a thick muscle located on the chest - ranging from the second to the sixth rib. The breast is composed of mammary glands, skin, and connective tissue. The mammary glands are modified sweat glands that develop during pregnancy, are maintained during lactating period and atrophy after this period ends. These mammary glands consist of fifteen to twenty lobes, that divide into lobules – the secretory part of the glands – that are located around the nipple. Each lobe is emptied by lactiferous ducts that dilate, forming lactiferous sinus that drain onto the nipple. Internally, the breast is composed by adipose and glandular tissue. In non-lactating women, the breast is mainly composed of fat while glandular tissue occupies a great part of the breast in lactating women. The breast is connected to the superimposed skin and to the pectoralis muscle through Suspensory (Cooper) ligaments that are fibrous bands of connective tissue. Vascularization is extensive across breast tissue - both from blood and lymph vessels – and the dermal blood capillaries and nerves are closer to the surface in the region that surrounds the nipple - the areola [10, 11].

Figure 1.3 depicts a general view of breast anatomy.

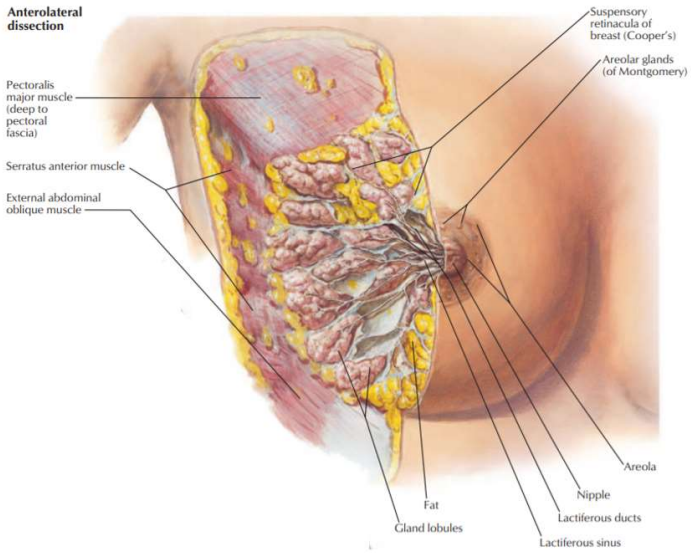

Figure 1.3- Breast Anatomy. Adapted from [12]

Concerning evolution of breast anatomy, at birth, only the main lactiferous ducts are fully developed. The mammary glands will only developed at its fullest during puberty, where the breast enlarges due to hormones like estrogen and progesterone, that stimulate the development of epithelial and connective tissue. Not only do the mammary glands develop during puberty, but during this time period there is also the deposition of adipose tissue, which contributes to breast enlargement. The breast will only complete its development during pregnancy, where the organ increases not only in volume but also in density, under the influence of various hormones. Besides breast enlargement, there are other anatomical modifications, which include vein dilation and a darkening in the pigment that constitutes the nipple and the areola. In addition, glandular tissue will start to occupy a great portion of the breast itself. Immediately after the end of the lactating period, the breast involutes, with glandular components and ducts starting to degenerate, resulting a diminished breast size. Finally, during menopause, ducts and glandular components will further deteriorate, making a menopausal breast to be primarily constituted by adipose tissue. However, over the years, there will be a decrease in this adipose tissue, which diminishes even further breast size and, moreover, the relaxation of the cooper ligaments may occur, resulting in breast ptosis [13].

1.3 – Breast Conditions

BC can be defined as a group of diseases where there is an uncontrolled cell division in breast tissue, usually beginning in the ducts or the lobules. There are two main types of BC: *In Situ* Carcinoma and Invasive Carcinoma.

As for In Situ Carcinoma, initially there were two classes: ductal carcinoma and lobular carcinoma; however, lobular carcinoma in situ is considered benign, although correlated with high risk BC development. Nonetheless, this condition does not have the potential to advance to an invasive stage. In contrast, Ductal in Situ Carcinoma is, generally, directly related to a development of invasive cancer. Nonetheless, while in an *in situ* stage, there is no proliferation of cells outside the location from where they were originated.

On the other hand, Invasive Carcinomas, which represents roughly 80% of all breast cancers, are diseases where cells emerge from the ducts or lobules where they first started to proliferate

uncontrollably. Invasive Carcinoma is a group of diseases with a vastly number of subtypes, depending on molecular characteristics [1].

Besides BC there are also benign conditions that can be related to a risk of developing malignancy. Some authors [14] divide these diseases in terms of risk of developing for BC:

- with no increased risk and minimal cell proliferation there are, for example, benign tumors like fibroadenomas, solitary papilloma, and sarcoidosis.

- with a small increase in risk and having cell proliferation without cell abnormalities there are diseases like ductal hyperplasia and sclerosing adenosis.

- with a moderate increase in BC with abnormal cell proliferation there are only two benign conditions: atypical ductal hyperplasia and atypical lobular hyperplasia.

1.4 – Breast Cancer Risk Factors

The question resides in what risk factors are not being considered when choosing the best screening option. Age is one of the best documented risk factors, with the incidence of BC being extremely low before the age of 30 and having a linearly increase until the age of 80 [15]. Body Mass Index has also been shown to be a potential risk factor for the development of BC but only in post-menopausal [15, 16]. Prior history of neoplastic or hyperplastic breast disease also presents itself as a risk factor for the development of BC. When it comes to family history, a woman who had a first-degree relative with BC when they were 50 years or older, is almost twice at risk of developing BC than a woman with no family history of BC [15]. Early menarche, late first full-term pregnancy and late menopause are three major risk factors for BC [17]. Normally, the earlier the age of the first menarche, the higher the cancer risk. The fact that both women with early menarche and later menopause are at higher risk of BC, can lead to the conclusion that prolonged exposure to estrogen is also a risk factor for this disease [17]. Longer duration of the breastfeeding period is associated with a diminished risk of BC, in comparison with women that had shorter breastfeeding periods.

Use of oral contraceptives also puts women at higher risk of developing BC [18]. As it was previously discussed, the existence of the BRCA1/BRCA2 mutated gene in women genotype puts them at higher risk of BC, compared to women who do not possess that gene [19].

Besides these risk factors, in 1976, John Wolfe, started studying the association between breast parenchyma patterns and BC. Wolfe showed that a prominent duct pattern helps to classify a woman as having higher risk than average for developing BC. Wolfe also stated that it is possible to predict which women will develop BC and which are less likely to develop it based only on the parenchymal pattern [20-23]. The studies conducted by Wolfe helped to define a classification of BC risk, based only on breast composition:

- N1, lowest risk, parenchyma composed primarily of fat, no ducts visible.
- P1, low risk, parenchyma chiefly fat with prominent ducts (< a quarter of the breast volume).
- P2, High risk, severe involvement of prominent duct pattern - occupying more than a quarter of the breast volume.
- DY, Highest risk. Severe involvement with dysplasia.

Many descriptors of these texture patterns have been documented. Mammographic density is one of those descriptors, normally represented numerically by percent mammographic density (%PD), that is also highly associated with an increased risk of BC [24-26]. In

fact, women with 60-70% PD are at 4 to 5 times higher risk than women with fatty breasts. Dense breasts are not only at higher risk of developing BC as are also more prone to aggressive tumors. As it can be seen, there are several risk factors that are not taken int account when deciding which methodology to pursue in terms of screening, imaging modality used and periodicity of the screening.

1.5 – Breast Cancer Screening

Screening programs all around the world use mammography, that can be acquired in a cranio-caudal (CC) and in a mediolateral-oblique (MLO) view, as a standard method for diagnosis, but although widely used, has both benefits and harms. The aim for an early detection of this disease started in the beginning of the 20th century with awareness campaigns, but a decrease in BC mortality was only observed when the first mammographic screening program started. On the bright side of mammography screening, life-threatening cancers will be detected early, improving prognosis, and consequently, decreasing risk of mortality. Studies point out that BC mortality rates, decreased at least 20% [27] thanks to an increase in mammographic screening - some studies even point out a reduction ranging from 30-50% [28]. Besides that, since cancer can be detected in an early stage, the available treatment can be less invasive and, consequently, have lower costs. The treatment will also be less intense, resulting in fewer time off of work, and, consequently, smaller money losses.

One of the problems associated with mammography is the rate of false positives. In Europe, the risk of having a false-positive result, for women in the range of 50-69 years having biennial screening, is about 20%. In what concerns to the United States of America, all screened women will experience one false-positive in their life. These false-positive results have an impact in women lives, especially in day-to-day well-being and in costs concerning healthcare. But the presence of false positive is not the only downside of mammography. A summary of the benefits and harms of mammography in 1000 women with a screening every two years showed that 200 of them will experience a false positive, 30 will have a biopsy due to the false positive result, 15 will be over-diagnosed and 3 will develop interval cancers. Interval Cancer is the name given to a cancer that appears between two consecutive mammograms. These interval cancers may have been developed between the two mammograms, however, around 35% of them were already present in the previous mammogram but were overlooked. This means that the patient received a false negative result that can occur because, in mammography, there is an overlap of tissue that can obscure the presence of cancers [29]. Since the population being screened is mainly composed of asymptomatic women, it is expected that with increased screening, it will also be seen an increase in cancer incidence. Life-threatening cancers will be detected early, improving prognosis, which is clearly a point in favor of mammography screening, however, cancers that would never be detected and that, in theory, were not harmful for the woman who presents it, will also be diagnosed. This is called overdiagnosis. Overdiagnosis leads to an ethical dilemma since there is a probability for the patient to live longer with cancer than with the treatment, and this decision-making process could lead to an increased anxiety state of the patient [29]. Another important aspect to consider, related to this type of screening, is the relation between mammography and dense breasts. The sensitivity of mammography decreases in women who have dense breasts (30-64% vs. 76-98% in women with fatty breasts) [27], which occurs because cancers have attenuation coefficients closer to dense tissue. Actually, a study from 1999 [30] showed that there was a significant trend between breast density and the appearance of false positives. Since it is known that breast density is an important risk factor for the development of BC, the fact that mammography does not perform so well in dense breasts should be of great concern. As

seen, there are multiple downsides to mammography and yet it continues to be the standardized screening method. However, in 2014, the Swiss Medical Board stated that the harms produced by these screening programs outweighed the benefits and, therefore, they recommended Switzerland to stop all the mammography screening programs [31].

Since mammography can mask cancer in dense breasts, it is expected that both women with dense breasts and their doctors want to go further in diagnostic techniques once mammography will not serve them. Ultrasound (US) presents itself as a complementary imaging method to mammography. Unlike other imaging techniques, US does not require ionizing radiation or contrast agents. Besides that, it is widely available, and it is inexpensive. When compared to mammography alone, the combination of these two modalities can increase cancer detection - with the downside of increasing the number of recalls. Studies have pointed out that US performs better in dense tissues when compared to mammography, detecting 2-4 additional cancers per 1000 examinations. Given all those factors, US has a tendency to grow in the screening picture. However, there are downsides with the use of US as a primary screening tool: limited capability to detect calcifications, long screening times and a need of highly trained technologists to perform the procedure. All of these previous factors may be the reason why Ultrasound is not as used as standard mammography in BC screening [27, 32].

To overcome the problem of overlapping tissue in mammography, a moving x-ray acquisition system around the breast can capture images from different angles that can then be reconstructed into a three-dimensional image - Tomosynthesis. Tomosynthesis can be used as a supplementary imaging technique to mammography. Studies point out that the use of these two modalities, combined, results in an increase in the rate of detection of additional cancers (30-35%). Besides being able to provide an improvement in cancer detection, tomosynthesis also performs better than mammography when characterizing the lesions and staging the tumor. Some trials point out that the use of Tomosynthesis and Mammography together not only increases cancer detection rate as also decreases false positive and recall rate [27]. However, there are some problems related to this type of image modality. The radiation used here ranges from 1 to 1.5 of the radiation used in mammography, so, using the two modalities together represents a great increase in radiation dose given to the patient. Apart from that, depending on breast thickness and the size of each slice, the time of reading could also be higher, when compared to standard mammography. The acquisition time with the use of both modalities can increase in 26% [33].

Even though mammograms are the standard procedure for screening women when it comes to BC, Magnetic Resonance Imaging (MRI) is widely used in women who are at higher risk. This happens, once again, because mammography is not so sensitive to dense breasts tissue. There are cases where women appear to have a local lesion (with mammography) and then are found to have a large-scale disease and often need mastectomy. This poor mammography sensitivity is also associated with younger women and BRCA1/BRCA2 gene carriers. MRI, on the other hand, has higher sensitivity and is not affected by breast density [34]. The question resides in what can be done differently to assess women at higher risk of BC. It is known that women who have the BRCA1/BRCA2 gene mutation develop BC at a younger age, that may not be covered by the screening program. Given that, a group of researchers [35] evaluated if the use of MRI surveillance in high-risk women improved survival status, and also compared the accuracy of BC detection for MRI + Mammography vs. Mammography alone. They concluded that not only the addition of the MRI screening to the mammography accurately detects cancers at a younger age as also shows a survival benefit between screened high-risk women and a non-screened group of high-risk women. Another group of researchers [36] stated that higher risk women should start mammographic

screening earlier and could even benefit from additional screening with other modalities - a contrast enhanced breast MRI. They went even further, declaring that all women should be evaluated for BC risk until the age of thirty, so that high-risk women are identified and could be medically accompanied, according to their condition.

1.6 – A deeper look into mammography

As it was already mentioned, despite its benefits and harms, mammography is the standard method when screening for BC and, for that reason, it will be the modality used in this project. For that reason, it is important to understand the basic concepts related to this imaging modality. Despite the problems related to breast density, this is still the imaging modality with highest effectiveness in detecting early-stage calcifications and has a relatively high sensitivity in detecting suspicious masses across all breast tissue. Mammography is a low-energy x-ray technique that aims to produce images with sufficiently high contrast to allow a fair distinction between healthy and cancerous tissue. Image can be captured either by using film screens or digital detectors, and this two equipment allow for a high spatial resolution. Given that, a mammography set is composed, in general, by an x-ray tube and receptors, an automatic exposure control device, a moving grid and a compression paddle to place the breast. The compression part is necessary because, during examination, breast compression decreases breast thickness, consequently allowing tissue to be better seen, decreasing exposure time, and thus decreasing radiation given to the patient. Besides that, the breast, being fixed, reduces the potential noise that could arise from movement. Still referring to detector types, nowadays, digital mammography is the standard type of mammography, but the main difference in image acquisition between screen-film and digital mammography is the type of detector used. Using a digital approach, postprocessing techniques become easier, the use of computer-aided detection systems gets simpler, and mammograms can be saved and transferred in an electronic fashion. Besides that, by removing screen-film from the apparatus, noise and artifacts that usually arise from film processing disappear.

Given the information above, it can be understood that mammography is a special X-ray technique. X-rays are produced by colliding high energetic electrons, accelerated due to electric potential difference between a cathode and an anode, into a target (anode). When they hit the target, their kinetic energy (that is due to the acceleration) is converted into X-rays. The radiographic image, as we know it, is a result of the different absorption of the X-rays by the tissues. When performing this and other modalities of medical imaging, it is important for the patients to understand if the benefits of radiation use outweigh the risks. In the specific case of mammography, and only concerning to radiation, the great risk that arises from the procedure is BC development 5 to 30 years after radiation exposure. However, this risk is inversely proportional to age, for example, a woman with 45 years has a lifetime risk of 1 in 100,000 of inducing fatal BC, as a result of a two-view mammogram; as the age progresses to 65 years, this risk drops below 0.3 also in 100,000. On the other hand, other statistics should be of greater concern: the probability of a BC to be present in a woman with 45 years is of 1 to 500, and the probability of that cancer to be fatal – without mammography screening – is of 1 in 4. When considering radiation problems related to mammography, the benefits seem to have preponderance over the risks.

Finally, it is important to understand the terminology used when looking and assessing a mammogram, since across the world, the medical field refers to mammographic findings in the same way. Doctors and physicians assess mammographic findings through a universal classification code, called American College of Radiology (*ACR) BI-RADS mammography categories,* but before going further into that, it is imperative to understand *what* can be found. There are two main finding in

abnormal breast tissue: masses and densities, and calcifications. For the first class, what differentiates a mass from a density is that a mass is a "space-occupying lesion" that can be seen in two different views, while a density is a possible mass, but is yet to be confirmed, because it can only be seen in one view. The assessment of these lesions takes into account its shape, margins and density. In terms of shape, the masses can be round, oval, or irregular, with the probability of malignancy increasing as the shape becomes more irregular. For margins, there are several classifications – circumscribed, obscured, microlobulated, indistinct, spiculated – and malignancy likelihood increases as the margins become more spiculated. Finally, concerning density, malignancy is highly related to a high-density measure. In what concerns to calcifications, there are two main groups: micro and macro-calcifications, with this last class not being of great concern, since most of the macro-calcifications are benign. Microcalcifications are small calcium deposits in the breast tissue and, when assessing them, both distributions across the breast and morphology should be considered. In terms of morphology, they can be: round and punctuate, being typically benign; amorphous or coarsely heterogeneous, being in an intermediate state between benignancy and malignancy; and fine pleomorphic or fine linear, having an higher probability of malignancy. Calcification distribution across the breast can also be an indicator of the lesion state. If calcifications are arranged in a cluster, there might be a probability or malignancy. However, if the lesions are organized in line, there is a greater likelihood of malignancy. The location where the calcifications occur might also point to lesion type, with calcifications occurring in blood vessels, fat or skin being normally benign [37-39].

In where it comes to the ACR categories, the assessment is given by evaluating the mammogram for the presence of suspicious lesions and their characteristics (presence of malignancy, for example). Based on that assessment, the classification into categories is as follows:

- Category 0: Incomplete; Carcinoma Risk unknown
- Category 1: Negative; 0% Carcinoma Risk
- Category 2: Benign Findings; 0% Carcinoma Risk
- Category 3: Probably benign finding; < 2% Carcinoma Risk
- Category 4A: Suspicious abnormality; 2-30% Carcinoma Risk
- Category 4B: Suspicious abnormality; 30-60% Carcinoma Risk
- Category 4C: Suspicious abnormality; 60-90% Carcinoma Risk
- Category 5: Highly suggestive of malignancy; 90-100% Carcinoma Risk

1.7 – Textural Analysis

Texture information is of great importance in BC risk prediction, as advanced by Wolfe. Texture is concerned with the spatial relationships between gray-levels and it can be considered as an important property of all surfaces, giving information about its architecture and relations with its neighborhood [29]. Although texture perception can be easily done through human eye, this application to computerized methods has some complexity associated. Textural patterns, based on the spatial variation of pixel intensity, are used to classify, for example, a surface as being fine or coarse. When assessing texture in mammograms, different sets of features can be considered and will be discussed further on.

1.7.1 – Intensity-based Features

There are features not directly related to texture, once they do not consider spatial relationships between pixels that are widely used in the field of BC risk assessment. These characteristics are

commonly called first-order statistical features, corresponding to properties of each pixel and not its relationship with the neighborhood. The retrieval of the said features is related to image's intensity histogram and the most commonly considered measures are: mean intensity pixel, maximum intensity pixel, minimum intensity pixel, 5% threshold – the gray-level that thresholds 5% of the area under the histogram, 30% threshold, 70% threshold, 95% threshold. Skewness, a measure of asymmetry of the histogram around the mean; and Kurtosis, a measure of the "tailedness" of a distribution, are also considered when it comes to the computation of intensity-based features [40]. From the referred thresholds, other metrics appear, like balance1, that is computed by dividing *a)* the subtraction between 95% threshold and the mean by *b)* the subtraction between the mean and the 5% threshold; and balance2, that has an analogous computation, using the 70% and the 30% threshold, respectively.

1.7.2 – Co-Occurrence Features

Co-Occurrence features, or Haralick features, are related to a special matrix called GLCM – Gray Level Co-Occurrence Matrix. It is the mathematical manipulation of this matrix that results in various features that are used to describe spatial relationships between pixels. It is, therefore, important to understand how the GLCM is constructed and how it can be manipulated in order for features to be extracted. Considering the image 'I' in Figure 1.4, it can be seen that the minimum pixel value is '1' and the maximum pixel value is '8'. Given that, the GLCM will be constructed by searching for all the possible pairs of intensity in the range of 1 to 8, and their co-occurrence will be registered. The search for this co-occurrence can be done in the directions of 0, 45, 90 and 135 degrees. As it can be perceived by Figure 1.4, the said search is being done in the 0-degree direction and, the co-occurrence (4,1) took place twice, which explains why the entry (4,1) in the GLCM has the number two. The same rationale is applied to all the pairs in the image. Even though the number of lines and rows range from the minimum intensity pixel value to the maximum intensity pixel value, these values are usually quantized to 256 gray-levels.

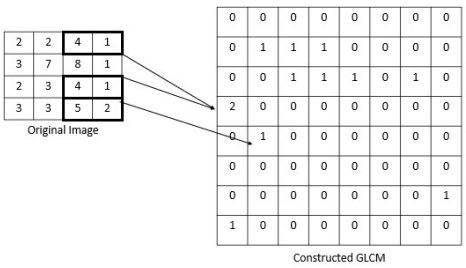

Figure 1.4 - GLCM construction, with the original image on the left and the constructed GLCM on the right.

A mathematical manipulation of the matrix allows the extraction of features like: Angular Second Moment, Contrast, Correlation, Sum of Squares, Inverse Difference Moment, Sum Average, Sum Variance, Sum Entropy, Entropy Difference, Variance Difference, Entropy Information Measure of Correlation I, Information Measure of Correlation II, Maximal Correlation Coefficient, Autocorrelation, Cluster Prominence, Cluster Shade, Dissimilarity and Maximum Probability and a formal definition of them was done by Haralick and can be found in [41]. These metrics are detailed in the Methods section.

1.7.3 – Run-Length Features

A Run-Length can be understood as a group, in an image, of consecutive pixels in a specific direction with the same gray value [42]. Run-Length Matrix (RLM) construction is done by registering the occurrence of sequences of a specific intensity in the image, in a given direction. As it happens in the GLCM, the search for a sequence can be done in the 0, 45, 90, or 135-degrees direction.

2	3	3	3
0	1	2	3
1	0	3	0
3	2	1	1

Original Image

	Run - Length				
Intensity	0 degrees	1	2	3	4
	0	3	0	0	0
	1	2	1	0	0
	2	3	0	0	0
	3	3	0	1	0

Constructed RLM

Figure 1.5 - RLM construction, with the original image on the left and the RLM on the right.

Each row of the matrix represents a pixel intensity, while each column represents the length of the registered runs. For example, considering Figure 1.5, looking for sequences of '3' in the "Original Image", in the direction of 0 degrees, it can be seen that the intensity '3' appears alone three times. This means that '3' has a Run-Length of one, three times, which is denoted by the entry (3,1) in the matrix. It can also be noted that '3' appears one time in a group of three, meaning that '3' has a Run-Length of three, one time, which is denoted by the entry (3,3) in the RLM. Given that, the conventionally extracted RLM features are: Short Run Emphasis, Long Run Emphasis, Grey Level Non-Uniformity, Run Percentage, Low Grey Level Run Emphasis, High Grey Level Run Emphasis. A formal definition of each of these entities can be found here [42].

1.7.4 – Additional Texture Features

The descriptors analyzed in the previous three subsections are the more frequently used and the ones that are better documented. However, other features can be used for BC risk assessment and a brief description of them will take place in this subsection. Fourier Analysis can be employed for the purpose of studying texture, concretely used in the task of characterizing breast parenchyma, with First Moment of the Power Spectrum (FMP) and Root Mean Square of the Power Spectrum (RMS) being commonly extracted features [43]. In the search for descriptors of a specific surface, in terms of texture, Local Binary Pattern (LBP) is of great importance. This approach defines a central pixel and analyzes the pixels from its neighborhood, attributing the value '1' to the neighbors with higher intensity than the central pixel and value '0' to pixels with lower intensity than the central pixel. In that way, concatenating the binary values in an anti-clockwise direction, each pattern will be defined by a single binary number that can be converted into decimal. This process is exemplified in Figure 1.6.

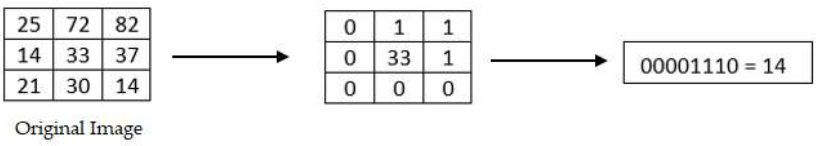

Figure 1.6 - LBP algorithm in action.

In line with LBP, Weber descriptors are also used to assess local parenchyma texture and breast tissue orientation [44], while the use of Gabor Filters, that have spatial responses similar to the ones presented by mammalians' vision, is related to whole breast parenchyma assessment. It should be noted that these filters are robust against image noise and low resolution [45]. Edge Frequency features, that compute the gradient and considers it a function of distance between pixels are also, although not very common, extracted for BC risk prediction [46].

Recurrent patterns that have the same architecture in different dimensions are called fractals, and fractal analysis, normally through a measure of the dimension of the fractals, can be performed to study different breast parenchyma patterns in mammography [46, 47].

Some authors use matrices apart from the GLCM to study features based in the relationships between gray-levels, more precisely, they use the Neighborhood Gray-Tone Difference Matrix (NGTDM). The method for the construction of this matrix is quite complex when compared to how GLCM and RLM are constructed, and it can be described as follows : The NGTDM will be a column matrix composed by the values *s(i)* with *i* being a gray-tone. First, considering *i(m,n)* a gray-tone of the image to be studied in the position *(m,n);* one should first calculate the average gray-tone of a neighborhood centered in *(m,n)*. Average (A_i) computation is given by equation 1.1.

$$A_i = \frac{1}{W-1}\left[\sum_{k=-d}^{d}\sum_{l=-d}^{d} i(m+k, n+l)\right] \quad (1.1)$$

with d being the neighborhood size and $W = (2d+1)^2$. Then, the entry of the NGTDM for each pixel can be computed through $s(i) = |i - A_i|$, if the number of pixels with intensity i in the neighborhood is different from zero, otherwise, $s(i) = 0$. The commonly extracted features from the NGTDM are Coarseness, Contrast, Busyness, Complexity, and Texture Strength, and their formal definition can be found elsewhere [48].

1.8 – Machine Learning

Considering the pursue of new diagnostic/prevention techniques, one term comes to mind: Machine Learning (ML). This a subdomain in the field of artificial intelligence that allows computers to learn by experience, using data [49]. ML models have a wide range of applications, with the most common ones being related to computer predictions, that are done based on statistical methodologies. A ML algorithm is routinely adapting itself based in received input, hence learning from experience, in order to accomplish a pre-defined goal. The process of architectural adaptation – training - is not only concerned in achieving the proposed task with the input data, but also in being capable of completing the same task, and producing the correct results, in unseen data – testing [50]. The question resides in what this proposed task might be: if a discrimination into previously defined labels is being made, for example, classifying a mammogram as belonging to a cancer/healthy patient, it is said that the task being done is a Supervised Learning task; however, if, on the other hand, the goal is clustering data, or finding hidden patterns in the data, then an Unsupervised Learning task is taking place [51]. Considering a supervised approach, the first step for developing a ML classifier is to create a dataset that is somehow designed with the goal of the project in mind. This dataset can be partitioned into a training set and a testing set. The first will be used to train the model – i.e., reshape is architecture while learning – and the last to assess model capability to achieve the classification goal

in unseen data.

Nonetheless, classifiers will not learn directly from raw input, so a task to be performed after dataset construction is to find descriptors, or features, in the dataset that, when extracted, will provide a correct classification. However, not all features that one can come up with have a good prediction performance, some of them might be redundant and others might even not have any discriminative capability. So, in order to grant that only important predictors are being considered, feature selection methodologies should be applied to the extracted feature set.

Feature selection can be done in a supervised, unsupervised, or semi-supervised/unsupervised way. The first uses labels in the data to select the most important features; on the other hand, unsupervised methodology is more ambitious, with no labels being used, resulting in a good way of discovering hidden meaningful information; finally, semi-supervised/unsupervised methodology has characteristics from both approaches, working on both labeled and unlabeled data. To search for the optimal set of features, one evaluation criterion should be chosen. There are four evaluation criterion classes: Filter, that examines each feature through its own intrinsic characteristics; Wrapper, that relies in a learning algorithm performance (accuracy) to choose what features to add or remove; Embedded, that as the name implies embeds the process of feature selection in learning algorithm development; and Hybrid that can be a combination of two previously referred criteria, a combination of two methods for the same criterion, or even two feature selection approaches. Other issues concerning feature selection methodology, like feature search direction (starting with a full/empty feature set, or other variations) or search strategy can be found in [52].

Once the optimal set of features is chosen, it is possible to give it as input to the learning algorithm, or classifier. This algorithm will learn, from the training set, to make distinctions between each previously defined class labels. Nonetheless, a classifier that only learns how to distinguish training examples is of no use, once the label for those cases is already known, it is therefore important that the trained classifier is able to generalize to unseen cases. This is where the concept of cross-validation enters: this is a technique that can be used to choose the best parameters for the model, once it evaluates how well does the classifier generalizes for unseen data, which can later be demonstrated by testing the model in an unseen test set. Cross-validation process consists in partitioning the training set into numerous sub-sets that are alternatively used for training and testing the classifier. The most common techniques are K-fold, Holdout and Leaveout cross validation. The first splits the data into k subsets of equal size, $k-1$ sets are used to train the model while the other is used to validate it, this process is repeated k times. Holdout validation divides the data into two subsets with a pre-defined ratio between them, train and testing are only performed once. Finally, Leaveout cross-validation will use the same approach as *k-fold*, but here k equals to the number of cases in the data. Since cross-validation performance is an indicator of the classifier generalization, model hyperparameters can be tuned using results from this validation technique. Finally, a testing phase can be performed using the previously defined testing set, in a classifier that was trained, validated and, consequently, has its hyperparameters optimized.

There are several metrics by which a classifier can be evaluated, like for example the number of correct guesses – accuracy. However, this metric might not be very useful, for example, if a dataset has 99 examples with label ‘0’ and 1 example with label ‘1’, an algorithm that just classifies every input with ‘0’ will have a 99% accuracy in the said set. However, one can argue that in fact, it is not a very good algorithm. A more interesting metric is the Area Under the Curve (AUC). AUC represents the area under the Relative Operating Characteristic (ROC) curve, which shows the performance of a specific classifier when the classification threshold is being modified. To trace this curve, two metrics are computed, True Positive Rate (TPR) – also known as sensitivity and recall - and False Positive

Rate (FPR) – also known as specificity. These metrics are calculated as shown in equations 1.2 and 1.3.

$$TPR = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (1.2)$$

$$FPR = \frac{False\ Positive}{False\ Positive + True\ Negative} \quad (1.3)$$

By displaying FPR in the *x* axis and TPR in the *y* axis for each classification threshold, ROC curve is traced, and AUC can be calculated. This area is going to be comprised between 0 and 1 and, the higher its value, the higher the performance achieved by the classifier. Considering the scenario of differentiating patients with and without a disease, a classifier that had no discriminative power would present an AUC of 0.5 with the ROC curve being a straight line. Any other line above this one would represent a classifier with higher discriminative power, like, for example, a classifier with an AUC of 0.8, meaning that there is an 80% chance of the classifier to correctly differentiate positive and negative cases. This comparison can be seen in Figure 1.7.

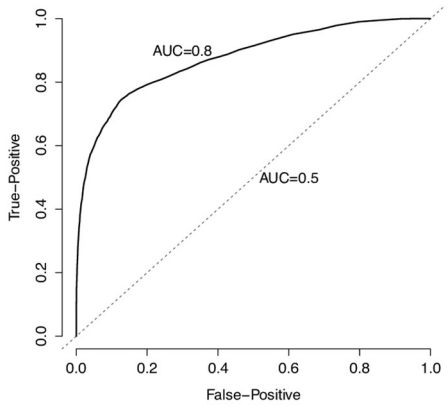

Figure 1.7- Comparison between an AUC of 0.5 and an AUC of 0.8. Adapted from [53]

Besides these metrics, a very famous score called F1 can also assess classifiers' performance. This metric, shown in equation 1.5, is computed through the conjugation between recall and precision (equation 1.4).

$$Precision = \frac{True\ Positive}{True\ Positive + Fa \quad Positive} \quad (1.4)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (1.5)$$

1.9 – Goal

This thesis aims for different goals: first, a differentiation between cancer (malignant) cases and healthy tissue (without lesions or with benign lesions) is aimed, through mammogram analysis, using ML methodologies. The main target of this task is to develop an algorithm that correctly classifies cancer and healthy patients without the need of pre-defining a Region of Interest, and that is not biased by image area, or breast density. A study on how noise disruption affects the classifier, and how filtering techniques can diminish the said disruption is also conducted. Besides that, the prediction capability of %PD is also going to be studied. The final use of the developed classifier is the development of an interactive application that allows the users to load the images and classify

them. Finally, in terms of risk, the goal is to analyze feature variation across years, using sequential mammograms, in order to predict if there is, or not, a risk of developing BC.

Across this chapter, clinical background about the breast and related diseases was given. Besides that, information about risk factors and concerning cancer screening was addressed. Finally, an introduction to texture characteristics and to ML was performed, before introducing the main goals of this project. Given that, the next chapter will review state-of-the art concepts, related to the methodologies used in this thesis, besides addressing research with similar goals to the ones presented in subsection 1.9.

2. State of The Art

In this section, state of the art processes, concerning the approaches used will be conducted. In subsection 2.1, image pre-processing methodologies, like image normalization, noise and filtering techniques are analyzed. Subsection 2.2 is concerned with the review of image registration techniques. Besides that, articles that aim for similar goals as this work – BC detection and risk assessment, will be analyzed in sections 2.3 and 2.4, respectively. The work present in section 2.4 is a summary of a peer-reviewed article that was an outcome of these thesis and that deeply reviews Machine and Deep Learning applications for BC risk prediction [54].

2.1 – Image Pre-processing

To better analyze images in order to discover patterns that indicate the presence of a tumor, or even in order to discover parenchymal characteristics that show a tendency for future cancer development, raw images need to be pre-processed. Usually, images may be acquired from machines with different characteristics, which could lead to pixel values that are incommensurable. Even images acquired with the same scanner, in two different points in time might present this relative difference – due to different acquisition parameters chosen by the technician. In order to overcome that and to grant that image comparison is done in a correct way, image normalization must be considered. There are several normalization techniques, and the straightforward ones will be explored, with the simpler one being the *Min-Max* approach. This normalization technique will consider all pixel values and transform their range into the [0,1] interval. Pixel distribution remains the same, the only thing that happens is that each pixel value is transformed by a scaling factor. *Decimal scaling,* another normalization technique, can be applied by considering a logarithmic scale for the distribution of the considered entities. The scaling can be done by dividing each pixel value by the overall maximum value between each distribution. For example, if there is one entity (image) with a distribution between 0 a 100 and another with a distribution between 0 and 1000, the new pixel values will be obtained by dividing each value by 1000. One of the most used normalization techniques is the *z-score*, which needs the mean and the standard deviation of the data to be normalized. These two values need to be known, if not, they will have to be estimated. Then, the new values after the normalizations are given by subtracting each pixel value by the overall data mean and divide the result by the data standard deviation. However, z-score normalization does not guarantee that different images will have the same numerical range after normalization. This z-score represents how many standard deviations is a datapoint away from the mean of that data. Finally, the *median and median absolute deviation normalization* can be performed, where each pixel value is subtracted by the mean and divided by the median of the absolute value of the subtraction between each pixel value and the absolute median. There are some problems related to each of this normalization algorithms; for example, the *Min-max* score is very sensitive to outliers, while on the other hand, the *decimal scaling* approach is limited, once it needs the data to vary by a logarithmic scale. As for *z-score* normalization, if the input data is not Gaussian, this type of procedure does not guarantee that the pixel distribution remains the same, besides the fact that numerical range of different images might not be the same after normalization; and for the last normalization procedure, the effectiveness is lower than the remaining because the median is not as strong as mean or standard deviation, statistically speaking. A deeper definition of these and more complex normalization techniques is given in [55].

Mammograms may present some artifacts that one might not want to consider for texture analysis. They may present, for example in the form of labels that must be removed, either by blurring or filling operations. With this in mind, morphological operations arise as natural solutions for this problem. Two of these operations are erosion and dilation. While in the first small holes are filled and some objects may become connected, in the second, as the name implies, objects' boundaries are eroded. For each of these operations a specific structuring element needs to be defined, which is going to control the manner by which the morphological operation occurs.

Both in dilation and erosion, the structuring element will slide with its origin across all pixel values in the image, with padding happening, if needed (zero in case of dilation, one in case of erosion). In dilation, if any of the superimposed structural element values match the image pixel values in the neighborhood, then the analyzed pixel becomes '1'. On the other hand, in erosion, for an analyzed pixel to be considered as '1', there has to be a fully match between the structuring element and the image. Results from an example concerning erosion and dilation are shown in Figure 2.1.

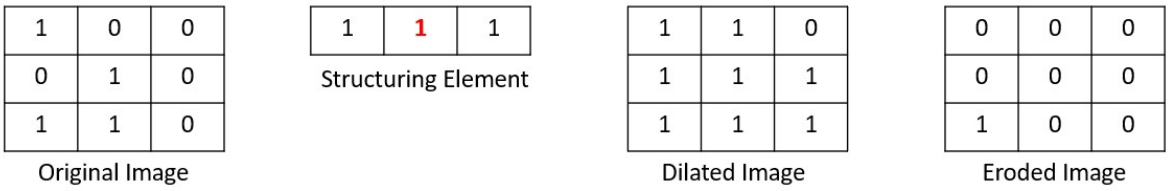

Figure 2.1 - Example of dilation and erosion, with a 1x3 structuring element having the middle '1' as its origin.

The combination of these two operations can be done in different ways, with the most used being opening and closing. The first allows boundary smoothing and protuberance elimination. This operation is achieved by performing an erosion followed by a dilation. Closing is, therefore, a dilation followed by an erosion and although may also smooth edges, it is more related to eliminating holes or fill gaps. Given the general depiction of these morphological image operation, many different algorithms can be written for several different goals, like boundary extraction, image filling, pruning and so on [56, 57].

To serve the purpose of this work, noise must be given to the image, and, after that, noise reductions techniques need to be applied. Considering that, it is important to understand the different types of noise that can be added to an image, and how that noise can be eliminated (or at least reduced).

One of the most common forms of noise is Amplifier or Gaussian Noise, which is commonly directly related to the image sensor. This is a form of white noise and arises from a random variation in the signal. This type of noise depends on the signal intensity, with each original pixel intensity being distorted by a small amount. If one would plot the amount of pixel variation due to noise versus its frequency of occurrence, a normal distribution of the noise would be see, hence the name gaussian.

Salt-and-Pepper or Impulse noise can be caused due to abrupt perturbations in image signal. This type of noise usually does not affect all pixel values, but only a small amount. This type of noise is usually related to poor lightning and dark perturbations. Images that present this type of noise generally present pixels with dark coloration alternated with original bright areas, hence the name salt-and-pepper. Errors in analog to digital conversion, dead pixels, and dust in the image acquisition system are common causes for the appearance of this noise.

Shot Noise is related to the variation in the number of photons that is sensed in a specific exposure level. This kind of noise admits a Poisson distribution, with each noise value at different pixels being independent from one another.

Finally, Speckle or Multiplicative noise, is a type of grainy noise that results in an increasing

mean grey level of the local area where it is being added. This is one of the most concerning types of noise once image interpretation becomes extremely hard when it is present.

A visual perception of different noise types can be seen in Figure 2.2.

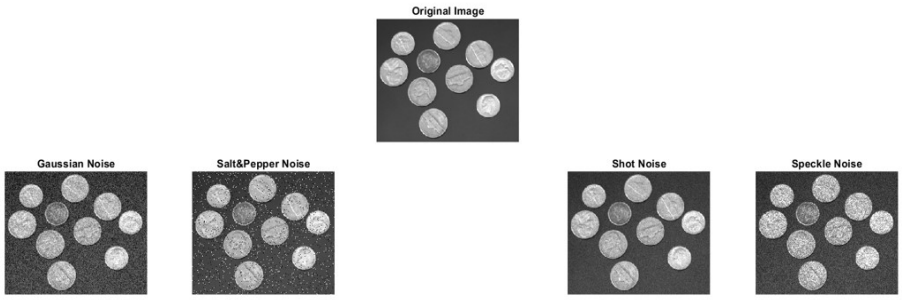

Figure 2.2 - Different Noise Type application to the same image.

When dealing with noisy images, filtering presents itself as an important step before processing the said image. There are several types of filters, starting with linear filters, these are very useful to remove noise but have also numerous downsides like resulting in blurred edges, image detail destruction and poor performance in noise that is signal dependent. Smoothing filters, that are a type of linear filter, commonly set each pixel value to the average of its neighborhood, which means that high intensity pixels are brought down by its surroundings, explaining why these filters tend to degrade image's edges. A linear filter which adapts itself to the local image variance, adaptive filter, tens to produce better outcomes. In this case, for example, if there is a high local variance, the filter performs less smoothing, once a higher variance might indicate the presence of edges. If the local variance is lower, then adaptive filters will provide more smoothing. Finally, nonlinear filters, like median filtering, provide a powerful tool for noise reduction with edge preservation. These filters are less sensitive to outliers, like the example of the higher intensity value in the average filter. Median filters is generally used to reduce the variation volume between proximal pixels [58].

2.2 – Image Registration

The alignment of medical images can be done with multiple purposes, like, for example, it can needed in order to assess mammogram characteristics for the same women across years. The alignment can be done in two different methods: feature-based techniques, and intensity-based. While the latter is dependent only on pixels intensity values for performing image registration, the first technique is more complex. As the name implies, feature-based registration relies on image feature extraction – edges, contours, surfaces – to perform the alignment [59]. Considering mammogram alignment, and having in mind that during time, breast changes its composition, beside the different breast positioning and compression that occurs at different acquisitions, a feature-based method that takes shape into account, is not a proper line of work to pursue. Given that, intensity-based methodologies will be considered. Besides choosing the methodology, one should also define an evaluation metric for the registration accuracy - the generally chosen metrics to evaluate registration performance are the commonly known measures of cross-correlation, sum of square differences, absolute difference, and mutual information (MI). Finally, a transformation type must be specified.

Image registration can be understood as a correspondence between two (medical) images, both structure and intensity-wise [60]. Given two images I_1 and I_2, their intensity matching can be done through equation 2.1.

$$I_2 = g(I_1(f(x,y)) \quad (2.1)$$

Considering *f(x,y)* as a spatial transformation applied to the images' coordinates, and *g* as an intensity transformation. Nonetheless, this intensity transformation might not even be needed, and one should only be concerned to finding the optimal spatial-coordinate transformation. The problem of finding the parameters that result in the best spatial transformation is the central part for guarantee a correct registration. This is a parametrical problem that can be implemented through equation 2.2.

$$I_2(x,y) = I_1(f(x,y) \quad (2.2)$$

There are several types of transformations that can be performed during registration but only the more common will be assessed, considering what it is allowed by the software used in the scope of this work. The simpler transformation is a *translation,* where each point (x_1, y_1) of the image will be shifted the same number of points/pixels in a given direction, resulting in point (x_2, y_2). This transformation can be expressed by equation 2.3.

$$T(x,y) = (x + a, y + b) \quad (2.3)$$

Rigid transformation allows not only translation, as also rotation and scale (however, a different subdivision can be made, was it would be seen further on). The general form of a Rigid Transformation has four parameters, a *t* vector accounting for the translation, an *s* factor accounting for scale, and a matrix *r* accounting for rotation. Points (x_1, y_1) are transformed into points (x_2, y_2) through equation 2.4.

$$\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} t_x \\ t_y \end{pmatrix} + s \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} \quad (2.4)$$

In Rigid Transformations, angles and relative lengths are preserved after registration, which happens because the matrix that is responsible for the rotation is orthogonal. The scale factor changes absolute lengths, but it occurs by the same scale in each direction. Concerning Affine transformation, the geometric modifications that are allowed are more complex, like the existence of shear transformation, and aspect ratio modifications. The general form of an affine transformation is given by equation 2.5.

$$\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} a_{13} \\ a_{23} \end{pmatrix} + \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} \quad (2.5)$$

Since the rotation matrix is no longer orthogonal, angles and lengths of the images are not the preserved after registration. Nonetheless, parallelism is maintained. As it was perceived more transformations can be performed, like shear or aspect ratio – the scale between the horizontal and vertical axes – changes. Their components in the field of affine transformations are given by:

$$Shear_x = \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix}, Shear_y = \begin{pmatrix} 1 & 0 \\ b & 1 \end{pmatrix}, Scale = \begin{pmatrix} s_x & 0 \\ 0 & s_y \end{pmatrix} \quad (2.6)$$

In Figure 2.3 it is possible to see a depiction of the different types of transformations that were discussed.

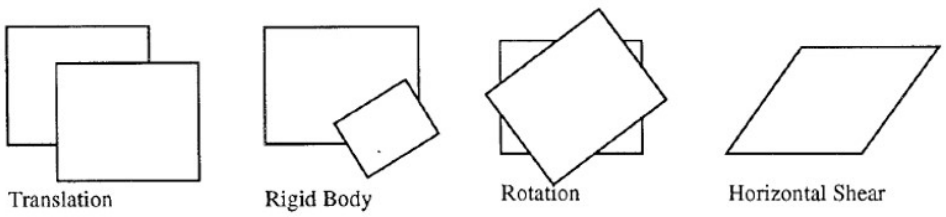

Figure 2.3 - Usual geometrical transformations [60]

A group of researchers [61] conducted a study to compare different methodologies for mammography registration. The data used consisted of pairs of images that were taken with a 23-month difference. The goal was to align these image pairs in a fashion that allowed the lesions that were present in images to be as close as possible. Four methods were used: *Nipple Alignment*, that consists in register the nipples through vertical and horizontal translation. This method has some drawbacks, once it is highly dependent on a good breast segmentation for nipple identification, and besides that, this method only works if there is a correct breast positioning in the mammography machine. *Center of Mass (CoM) Alignment* methodology is very similar to the last one, but instead of registering the nipple, it registers the CoM, therefore, it has the same limitations. MI was another used method, that is very popular even in multi-modal registration, once it is not dependent on breast positioning because it uses breast internal characteristics. This methodology allows translation, rotation, scaling and vertical shearing transformations. Nonetheless, the fact that scaling is allowed might be a disadvantage once this procedure might use scaling to compensate for tissue variation (loss or gain) across mammograms. Finally, a *Warping* method was used, which used an automatically determined set of points in breast tissue and pectorales muscle to make the registration. The results obtained by these authors proved that MI outperformed all the other methods used. Yet, the authors point out that background elimination might be an important step to improve the registration.

In a more recent approach, Famouri, Morra and Lamberti, [62] proposed a technique for registering different mammography views. As it was perceived, tissue overlap might obscure the presence of lesions in mammograms, or on the other hand, may cause a false positive. These authors consider, for that reason, important to combine information from both Cranio-Caudal and Mediolateral-Oblique views. In this research, a convolutional neural network is proposed in order to find the best possible affine transformation, so that the mean squared error between the two registered views is minimal. Since breast tissue is very complex, the authors proposed a supervised methodology to further improve registration. This is done by using lesion locations annotation as guidance for a good alignment.

Other research groups were conducted in the area of breast imaging registration with both intensity and feature-based methodology. A summary of some of these studies, the features, transformations, metrics and optimization methods used is given by [63].

2.3 – Automatic Breast Cancer detection using Mammography

In order to decide what approach to pursue in this work, it was important to first assess what was already done in the field of BC detection, and BC risk prediction, using not only mammograms but also the previously referred textural features. Concerning BC detection, a numerous amount of articles were screened and only the more relevant are going to be analyzed.

A group of authors [64] aimed to detect breast masses using an algorithm called Support Vector Machine (SVM). The pipeline of their work started by image preprocessing, with a Region Of Interest (ROI) definition being made in the lesion location. Besides ROI definition, lesion contrast was

also enhanced. Overall digital enhancement was also performed, using the Discrete Fourier Transform. Finally, from the pre-processed image, a segmentation procedure was performed, and the result was an image that contained only the mass, with all the remaining pixels "turned off". Once image was segmented, feature extraction was carried out, using mostly GLCM features. An image classification was then executed using the SVM algorithm, to perceive if the mass was benign or malignant. Overall accuracy of their model was of 86.84%, with sensitivity and specificity achieving values of 92.30% and 62.50%, respectively.

Another research group [65] aimed to, once again using Support Vector Machines, detect the presence of BC. Noise reduction and image contrast adjusting was done using a technique called Limited Adaptive Histogram Equalization. Besides that, a ROI definition was done in the lesion location. GLCM features were extracted from this pre-processed ROI in four directions and using two different distances between pixels. As for the classification phase, the authors aimed for a different approach from what is going to be performed in this work: instead of considering benign lesions and healthy patients as part of the same group, two different classification tasks were made. The first task aimed to differentiate normal and abnormal tissue, while the second and more difficult task had the goal to discriminate mammograms as being benign or malignant. For the first stage of the work, the differentiation between regions of interest that contained lesions and regions of interest of normal tissue, the system correctly classified all instances. For the differentiation between benign and malignant lesions, the systems achieved an accuracy slightly higher than the previous analyzed research, with a value of approximately 92%. Sensitivity and Specificity had values of 91.3% and 94.4% respectively.

Li [66] and his colleagues aimed to verify if adding information about the contralateral breast, instead of using only information from the lesion ROI, could improve the differentiation between benign and malignant lesions. Firstly, in order to computationally assess breast lesion, features that concerned size, shape, margin and spiculation of the tumor were computed. Besides that, common texture features were retrieved. After that, texture features were computed in the contralateral breast, that had no abnormality, recurring to a ROI placed in the region immediately behind the nipple. In each ROI, features that concerned to co-occurrence, fractal dimensions, edge-frequency and Fourier analysis were retrieved. After feature extraction, feature selection methodology was applied in order to select the best, non-redundant, set of features. The algorithm used for classification was a Bayesian Neural Network, and this task was done twice: the first using only the features retrieved from the lesion ROI; and a second time combining information from both breast tumor lesion and also the contralateral breast. Algorithm's performance when using the tumor features alone, presented an AUC of 0.79 ± 0.03 for the task of discriminating malignant and benign lesions. For the same task, when adding information from the contralateral breast, this metric value raised to 0.84 ± 0.03. The authors also used the classifier to check how the algorithm would perform if only the contralateral images were given as input, which result in a poor evaluation, having an AUC of 0.67 ± 0.04. This paper provided then a statistically significant information that the adding of contralateral, healthy, parenchymal measures, positively affects discrimination between benign and malignant images.

The aiming for a better differentiation between malignant and benign lesions lead a team of investigators to use an established and high-performance Decision Tree algorithm – C5.0 DT – to perform this task [67]. In order to improve image and lesion (calcification) contrast, a noise-reducing technique using a low-passed filter, was done by these authors. ROI selection was conducted in the region that contained the lesion. The extracted features retrieved from each selected region were related to both co-occurrence matrix and Run-Length matrix. The use of this features for the differentiation between the malignant and benign images present in the dataset resulted in a maximum accuracy of 96.7%. When assessing the AUC, this approach presents a value of 0.995, which is a very impressive result. The main contribution of this study was a proof of concept that the C.50DT

algorithm is a good tool to be used when performing malignant/benign differentiation.

Supporting BC Diagnosis through textural analysis of mammograms was also aimed by a student of the Institute of Engineering, in Polytechnic of Porto, with the differentiation between malignant and benign lesions being made, as also the distinction between normal and diseased tissue [68]. The used images had normal healthy breasts, and breast with some type of lesion - masses, calcifications, distortions, and asymmetries. Each mammogram was firstly smoothed and enhanced using low-pass and top-hat filtering. Feature extraction was conducted in a manually selected ROI that contained the lesion, with co-occurrence and Run-Length features being considered at various angles. Features were extracted twice from each image. One time in the ROI that contained the lesion, and other from a healthy/normal zone with the same size as the pre-defined ROI. After feature extraction and before classification, feature selection was performed in order to reduce feature dimensionality. For classification purposes, five different classifiers were tested: Naïve Bayes, SVM, k Nearest Neighbor (k-NN), J48 and Random Forest; and different studies were performed. First, the authors aimed to classify each image into benign and malignant, using only the Run-length features. The best result was obtained for Random Forest, with an AUC of 0.831. In the process of distinguishing normal and diseased tissue, the authors present results of a 100% accuracy. When trying to discriminate between normal, microcalcification and masses, although healthy tissue is always correctly identified, all the classifiers tend to fail when trying to identify masses, which may indicate that this type of lesions are more difficult to be correctly classified.

2.4 – Breast Cancer Risk Prediction

Before going further into parenchymal texture analysis for risk assessment through mammograms, it is important to understand how BC risk is currently addressed in clinical practice. There are three main models by which this is done.

2.4.1 – Gail Model

This model considers several epidemiologic risk factors in order to compute a risk score (a likelihood) for the future development of BC. The calculation is done taking into account: age at the first menarche, number of breast biopsies performed, age at the first live birth and first-degree relatives that had developed BC. This model considers a correction factor when dealing with women that had been diagnosed with atypical hyperplasia. Although this is the most used model, it fails to contemplate the history of lobular carcinoma in situ (LCIS). Other of the major flaws of this model is that although it examines the number of first-degree relatives with the disease, it does not take into account the age at each of these relatives were diagnosed. Besides that, the model fail to consider history of BC in the paternal lineage. Gail Model is the most accepted technique for risk assessment and, based on the results of this model, preventive treatment can be administered, with the Food and Drug Administration (FDA) allowing the use of Tamoxifen in women with a risk over 5 years higher than 1.7%.

2.4.2 – BRCAPRO model

This is a Bayesian model that aimed to provide estimates of the presence of the BRCA1/2 genes in a family, and this likelihood is used to compute the BC risk. To achieve these results, the model used the mutation frequencies, cancer penetrance in mutation carries, cancer statues of first-degree and second-degree relative, and their ages. The fact that the information from both affected and

unaffected relatives is taken into account may be an advantage of this model. Nonetheless, risk factors that are not hereditary cannot be considered for this model, which is a major disadvantage.

2.4.3 – Cuzick-Tyrer model

The approach proposed here was able to combine family history, measures of endogenous exposure to estrogen and benign breast disease history into a single model. This model allows to consider not only the BRCA1/2 gene but also other genes with lower penetrance. Cuzick-Tyrer model is able to compensate for the flaws of other models, once it considers, for example, the history of LCIS. Besides that, the age of the first- and second-degree relatives when they got diagnosed is also considered, overcoming some downsides of Gail Model. As it happens to the last model, this model will compute the likelihood of having some the possible mutations and, from that, infers the risk of developing BC, which possibly explains why this model is not as used as Gail Model, which directly computes a risk probability.

2.4.4 – Texture Analysis with Machine Learning

Textural features and Parenchymal tissue analysis are also used for purposes directly related to BC risk assessment.

When analyzing different approaches with this goal in mind, it is important to make a reference to regions of interest in mammographic feature extraction. Li *et al*., [43] studied in 2004 the effect of ROI size and location for feature extraction in BC risk analysis. The researchers defined five different ROI locations, as it is depicted in the picture below (Figure 2.4), by the letters A,B, C, D and E:

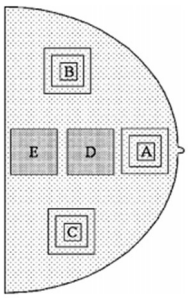

Figure 2.4 - Different Locations for ROI evaluation [43]

They wanted to see how both the location of the chosen ROI and its dimensions affected the analysis of risk, by evaluating the performance of the same set of features in differentiating high-risk and low-risk women. Women who had the BRCA1/2 mutation were considered high-risk and women who had a risk lower than 10% for developing BC in their lifetime - calculated by Gail's model - were considered for the low-risk group.

Features from the previously described groups were extracted from each ROI and a feature selection methodology was employed. The selected features were RMS, FMP, Contrast and Fractal Dimension. The discriminatory capacity was assessed for each individual features and for the selected features merged together. For the size analysis, larger ROIs presented better discriminative capacity (AUC) – 0.68 to 0.83 for individual features and 0.92 for the merged selected subset - than smaller ROIs, although without statistical significance. When it comes to ROI location, a comparison of the merged feature set discriminatory capacity across the five regions was made. The AUC for regions A,B,C, D, and E with size 256x256 was, respectively: 0.92, 0.78, 0.69, 0.84 and 0.79. The significance between area A and all the others was tested and showed statistical value. This research showed that

the discriminatory performance was substantially and significantly higher in region A, immediately behind the nipple, and moving the ROI from there to any other location, resulted in a statistically significant decrease in the performance. This is the reason why most of the researchers use the region immediately behind the nipple (retro-areolar region) as the ROI for feature extraction

However, in 2015 , Zheng *et al.*, [69] advocated that the use of a single ROI for feature extraction and BC analysis was not the more appropriate methodology. They stated that a single region in the breast may not be capable of considering all the heterogeneity present in breast texture. Their opinion is that texture analysis should be done with features being extracted from different structural elements across all the breast tissue. The hypothesis presented is, then, that the relation between texture analysis and BC analysis can be better analyzed if the retrieved features can better characterize tissue heterogeneity.

Given what was said about their hypothesis, the authors presented a method for computing the features across all the breasts - they called it the Lattice-Based Strategy. The approach consists of displaying a grid across the image and using a structural element - square - at each intersection node of the grid. Before displaying the grid, a preprocessing technique needs to take place. As it can be seen in Figure 2.5, there is a dotted yellow line outlining the breast - this was achieved by identifying the interface breast-air with thresholding methods. Besides that, the boundary between breast and pectoralis muscle was also delineated since the features need to be calculated only inside the breast. This breast segmentation is, therefore, an important step in image pre-processing for risk analysis.

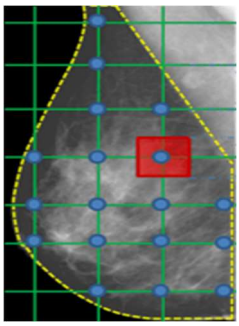

Figure 2.5 - ROI definition method proposed by [69]

PD% was also considered in this approach. In order to calculate this value, an automated method for breast partitioning is used. Once the breast is divided into subregions with similar image properties, an SVM algorithm classifies each region as being composed primarily of fat or dense tissue. Then, PD% is simply calculated by considering which is the percentage of the whole breast that is composed by dense tissue.

The remaining textural features will be computed in a window (seen in red) at each intersection point of the grid. Using this method, the whole parenchyma is characterized, since the features are extracted at different breast locations and not only in one ROI. Feature extraction and further selection was conducted and, in general, GLCM, RL, Histogram and Fractal dimension are more commonly chosen than LBP. When combining these texture features, the predictors outperformed the discriminatory capacity of PD% and no significant improvement was noted when PD% was added to the set of features. The lattice approach presented a discriminative capacity (AUC) of 0.85, in the task of differentiating high-risk women (using contra-lateral images of women with proven biopsy cancer) and controls. When using the exact same set of features, the single ROI methodology presented an AUC of 0.60 in the retro-areolar area and an AUC of 0.74 in the central breast area. The authors, when comparing the AUC values, concluded that their lattice approach

significantly outperformed the single ROI approach and validated that this outperformance is maintained when PD% is added to the set of features.

Another interesting research is the one done by Tan *et al*., [70] where the authors aimed to evaluate the viability of predicting BC risk in women after they had a negative mammogram. Given a sample of screened women, each woman was included if had had two consecutive mammograms acquired in the authors facilities and if the first mammography was negative - being defined as BI-RADS 1 or BI-RADS 2. A dataset was then created with the accepted women. Each case was composed by two mammograms - defined as 'prior' and 'current' evaluations - and, based on the current evaluation, the dataset was divided into three subgroups.

The first was composed of women who had positive results, confirmed with other evaluation methods. The second subgroup consisted of women who had abnormalities in their mammograms, were recalled but then the lesions proved to be benign. Finally, the third subgroup includes women with negative mammograms and that were not recalled. In the study, the authors used all the 'prior' evaluations to assess BC risk in the 'current' evaluation. It is important to add that, for each dataset case, age, family history of BC and the density rating by the BI-RADS scale were the epidemiologic risk factors considered. For feature analysis purposes, the authors segment the breast and extract the features in the segmented areas. Although these features are extracted, they will not directly be used for risk assessment, in opposite, what is done is that, based on the extracted features, features that characterize bilateral differences between mammograms laterality will be computed - Asymmetry Features. A set of 180 asymmetry features were calculated and, adding the epidemiologic data, a final set of 183 features was considered. Feature selection was performed, and a classifier was trained and tested with the referred dataset. The algorithm outputted a score ranging from 0 to 1. The higher the score, the higher the probability of the women to have an 'image-detectable' cancer in the next screening. Considering the first and third subgroup of the dataset and concerning a classification for Negative or Positive scenarios in the "current" mammogram, the AUC of the classifier was of 0.725. The study provided optimistic results in its aim of developing a new risk stratification tool that can help physicians to decide the periodicity of screening to each patient.

Li *et al*. published, in 2012, [19] a paper that aimed to prove the robustness of textural extracted features in the task of distinguishing between two groups: high risk and low risk women. A different approach was done here, since the high-risk group was not only composed of women with the BRCA1/2 gene mutation but also of women with unilateral cancer. Once again, for feature extraction, the region immediately behind the nipple was considered as ROI. Although the authors do not specifically state what features were selected, it is said that they described homogeneity, coarseness, non-linearity and randomness of the image texture, which can be linked to what was presented in section 1.10.1. When distinguishing BRAC1/2 high risk from low-risk women, the classifier had an AUC of 0.82. When the task was distinguishing the unilateral cancer group from the low risk, AUC was equal to 0.73. Finally, when merging these two subgroups together, the classifier presented an AUC of 0.75. Once again, these results proved that textural features represent an important tool in BC risk assessment, once there is a significant difference between high-risk and low-risk women.

In 2020, Gandomkar *et al.* [71] studied the combined effect of both textural analysis and epidemiologic factors for BC risk prediction in Chinese women. One of the differences between this research and others conducted with the same aim, is the epidemiologic features. Nonetheless, these authors used contralateral images from woman who presented cancer for the high-risk group, which is a significant difference, when compared to similar research. While the retrieval of textural features is practically the same as used in other articles- GLCM and Fractal Dimension- , the epidemiologic

features have a much higher range and are not restricted to the common: age and family history, showing promising results. Here, height, weight, age of menarche, breastfeeding duration and many other risk factors discussed in section 1.8 are considered. The authors hint that the combined effect of epidemiologic factors and texture features results in better prediction outcomes, as it was shown by other studies . Actually, in this specific study, the AUC for the classification, using Decision Trees, was of 0.88.

As it was perceived, there is extensive research done in the field of risk stratification. However, much of the developed work is related to the classification between low-risk and high-risk women. The proposed work aims in a different direction - to give a risk score for each specific patient, in order to help physicians to choose the best diagnostic procedure for the said patient. Although Tan's research has some convergence points with the goal of this project, some limitations of the developed work must be taken into account: The fact that the dataset used was produced in laboratory does not reflect the ratio between positive and negative cases in common BC screening programs; the methods used for validation may have resulted in bias and, the fact that the same portion of the dataset was used both for features selection and to evaluate the classifier accuracy may also have resulted in some bias in the process of optimizing the algorithm. Besides that, only asymmetry features were computed, which could lead to some masking effects of the effective texture of the parenchyma.

The work aimed here was more related to the works of Tan and Gandomkar than the others. While Tan uses asymmetry features between each breast to assess risk, and Gandomkar relies on contralateral imaging to consider women as part of the high-risk group, in this project, mammograms acquired several years prior to a cancer diagnosis were used to assess cancer risk, in a fashion described in further sections.

Once a theoretical background about the methodologies to be used was given, and after information about its applications was analyzed, the next chapter will address the dataset used and the methods applied to achieve the proposed goals.

3. Materials and Methods

There are six different subsections in section 3. The first gives an overview about the data used. Subsection 3.2 is concerned with the different methodologies applied in pre-processing the images, from background removal to image noising and denoising, passing through image normalization. In subsection 3.3, feature extraction is explained. The methods by which feature selection occurred and the posterior classification approaches, using the selected features, are explained in subsections 3.4 and 3.5, respectively. Finally, in section 3.6, risk approaches are explored.

3.1- Dataset

For the first part of the work, initially, images from both the Curated Breast Imaging Subset of Digital Database for Screening Mammography (*CBIS-DDSM)* public dataset and from *Hospital da Luz* were used to constitute a novel dataset, composed by 232 images. *Hospital da Luz* contributed with 84 images – belonging to 14 women -, all controls. The remaining part of the dataset was composed by *CBIS-DDSM* images, having 75 benign lesions and 73 malignant cancer cases. Hence, in the dataset there are nearly two controls for each cancer case.

Hospital da Luz images were collected from 2007 to 2017; Siemens Mammomat Novation DR was used in the acquisition of images from 2007 to 2008, and Siemens Mammomat Inspiration for the remaining years. X-ray tube current ranged from 97 to 168 mA, and the voltage used was between 27 and 32kV, with the parameters being automatically fixed for each subject.

The *CBIS-DDSM* dataset is an updated version of the Digital Database for Screening Mammography, containing more than two thousand images of digitized film mammography. From this dataset, malignant lesions, and benign lesions that did not became malignant were used, with this last group being considered as control. So, while the "cancer group" was composed only by images with malignant lesions, the control group had images with healthy tissue and images with benign lesions.

Dataset division led to the creation of a training (N=162) and a testing set (N=70). A balance between calcification and mass lesions, as between different density classes amongst cases was aimed when constructing the training set. Besides that, it was noted that using a single test set to assess classifiers' performance could result in some bias. For that reason, the firstly created test set was not considered *per se* and four new test sets were created. In order to do that, the images that were part of the original test set, along with all the images of the *CBIS-DDSM* dataset that were not considered to be used in this thesis, were merged in an image set from where it were randomly drawn four sub-sets of 70 images. This process resulted in four randomly created test sets, with the only constraints being: a) a balance between controls and cancer cases, and b) for malignant lesions, a balance between masses and calcifications. Dataset creation, a its division into training and testing set can be seen in figures 3.1 and 3.2, respectively.

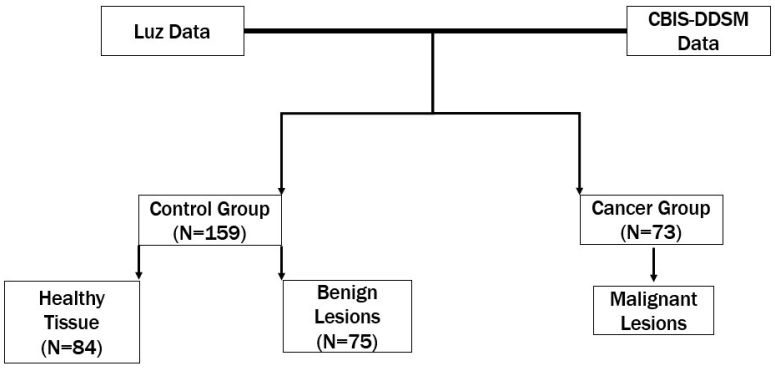

Figure 3.1 - Dataset creation

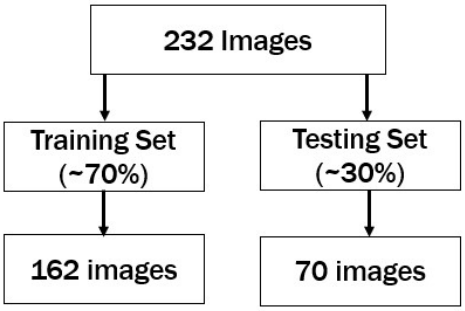

Figure 3.2 - Division into training and testing set

Information about lesion type, shape, margins, and breast density of the *CBIS-DDSM* dataset is available in [72]. Image sizes varied from patient to patient both in height and width.

For Risk Analysis, only *Hospital da Luz* images were used once, for risk assessment, information about future development of BC and/or multiple images for each woman was needed. Therefore, 11 different cases were considered. Four women had cancer diagnosed in the year immediately after the last image in the dataset was taken and, for that reason, were considered part of the "high-risk" group. The remaining seven women composed the "low-risk" group.

3.2 -Pre-Processing

3.2.1 – Background Subtraction and Image Normalization

In order to understand what was done in this preliminary step, one should look to a general depiction of the images in the dataset. Mammograms are usually composed by breast tissue, muscle tissue, surrounding background and labels that indicate view and laterality of each image.

A general depiction of the mammograms present in the dataset can be seen in Figure 3.1.

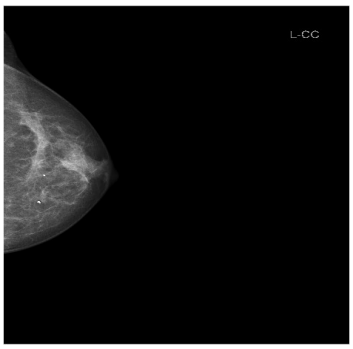

Figure 3.3 - Image from dataset

Since the aim of this first stage of work is to assess and differentiate cancer and non-cancer cases through texture analysis, the only part of the image that matters is referent to breast tissue. The presence of muscle tissue could be a problem, however, that should only be noticeable in MLO views, and, for that reason, this problem can be countered by using CC views. As a matter of fact, only cranio-caudal views with left laterality were used in this work.

Therefore, in order to assess only breast tissue, background removal should be performed.

In general terms, image pre-processing started with a definition of the boundaries that separate breast tissue from background, and consequential removal of extra-background. As it can be seen in Figure 3.1, besides images' labels and the breast itself, the image appears in black, which means that all background pixels have an intensity close to '0'. Background removal was performed by looking, from left to right direction, for the first column that had all values equal to zero. This column approximates the nipple *x*-coordinate. To guarantee the best possible background removal, an analogous search was conducted in the top-down and down-top direction, but here by looking for the first row in which at least one element was different than zero. After finding the columns and rows that separated breast tissue from background, all the pixels that were not within the limits defined by them were eliminated.

To achieve this goal, two different problems needed to be addressed first: one must assure that all background pixels are in fact zero, which may not happen due to variations in X-ray exposure during screening and, besides that, to avoid problems with images' labels when making the horizontal and vertical search, one must first remove them from the image.

So, actually, the first step towards background removal is label elimination which starts with image binarization, that it is achieved through Otsu's method. This is an automatic and unsupervised method to find the best possible threshold to segment the image between background and foreground. The said threshold is found by maximizing a discrimination criterion that measures the separability between the gray-levels of the two classes – background and foreground.

After performing binarization, the morphological operation of closing was done. Closing, as explained, is an operation that consists in image erosion after performing dilation in that image, with the same structural element. In dilation, each pixel gets as new intensity the highest pixel value in the neighborhood, while in erosion the opposite happens. A disk of radius 50 was used as structural element.

The application of this morphological operation resulted in the labels appearing as a hole surrounded by black pixels. After that, the complement of the resulting image was computed, which led the labels to appear as a black hole surrounded by white pixels. Following that step and using *built-in* MALTAB functions, the gap present in the image was filled. This sequential process can be observed in Figure 3.2.

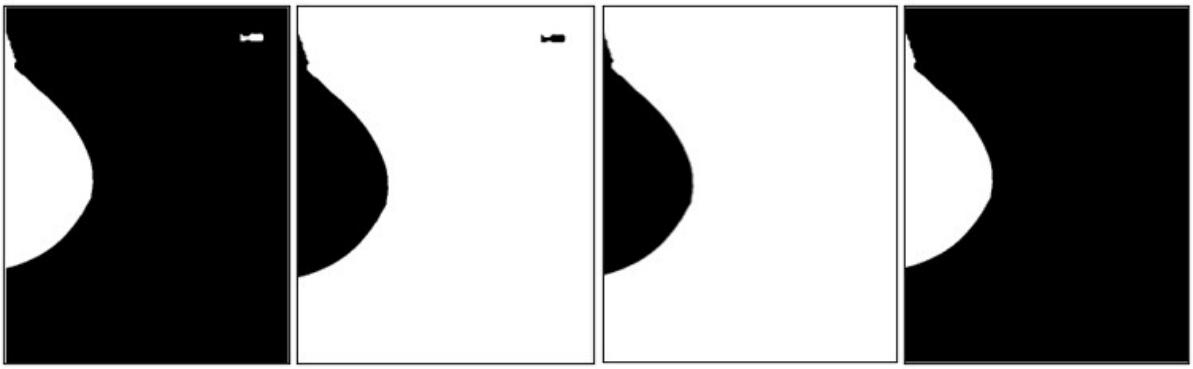

Figure 3.4 - Label Removal Process, starting with the closing operation results, moving to the complement image, followed by the results of the filling operation. The final image represents the binary image used for vertical and horizontal search.

Once label removal was done, the image was complemented back, resulting in the rightest image in Figure 3.2. The previously described search in horizontal and vertical directions was performed in this image. With the obtained results, image cropping took place by considering only the pixels within the limits defined by the rows and columns that were found, hence removing background that does not contribute for texture evaluation. The result of this process can be seen in Figure 3.3.

Figure 3.5- Result of Background removal process

To account for differences in overall pixel values, either because images were obtained through different machinery, or simply because they come from different datasets, a normalization step must be considered. In this specific case, normalization can be understood as a pixel value-based adjustment that guarantees that all images are in the same intensity scale. For this work's purpose, pixel values were scaled into the range of 0 to 1. Considering I as the image to be scaled, I_{min} the minimum pixel intensity value, and I_{max} the maximum pixel intensity value, image normalization process could be done according to equation 3.1.

$$I_{normalized} = \frac{I - I_{min}}{I_{max} - I_{min}} \quad (3.1)$$

3.2.2 – Noise Adding and Filtering Technique

In order to study how noise affects the classifier, image degradation should be done to each dataset image, so that it simulates, for example, dose reduction. It can be noted that most of the noise that appears in digital images –resulting from the image acquisition system - can be simulated through gaussian noise [73]. Authors [74] even point out that once gaussian noise models are so easily mathematically manipulated they are widely used in a great range of applications, even if they are only "marginally applicable". This fact gives confidence to proceed with this methodology for noise simulation.

Two different factors needed to be defined when adding gaussian noise: mean and variance of the distribution. For the first variable, a mean of 0 was defined, as it is common practice for normally distributed models; in where it comes to variance, the approach was significantly different.

Instead of giving a fixed variance value to the distribution and apply it to the whole image, the rationale was to make the variance dependent on the image pixel values, meaning that each pixel will receive noise from a specific distribution. To do that, a local variance value needs to be defined in order to create a Local Variance Matrix, that will define the variance of the distribution at each pixel location.

The process of creating this matrix for a local variance of 0.01 can be seen in Figure 3.4. For example, pixels with intensity 3 in the original image, will receive noise through a distribution of mean 0 and variance 0.004.

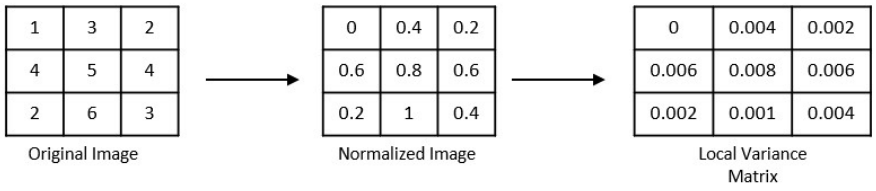

Figure 3.6- Local Variance Matrix definition for a local variance of 0.01, applied to the normalized image through the Min-Max approach.

In this work, noise adding was performed for local variances of 0.01, 0.005, 0.0025 and 0.001, simulating different noise magnitudes, where the higher the local variance, the higher the noise values added to the image. This procedure resulted in four noisy variations of each image, each with different noise magnitudes.

When adding noise values to each pixel of the normalized image, it should be expected that some intensity values will become higher than one and therefore not falling within the previously defined range. Given that, after noise adding, image normalization is performed once again.

A comparison between the original image, and a noisy image with magnitude of 0.01 can be seen in Figure 3.5.

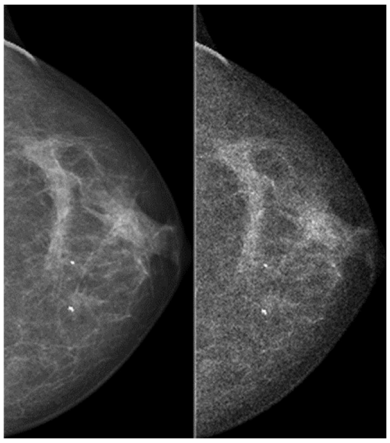

Figure 3.7- Original Image, on the left, and an Image with noise magnitude of 0.01, on the right.

Filtering routines could help to counter noise appearance in the images and, consequently, to recover information that might be lost with dose reduction or with outdated machinery. It must be noted that the filter to be used should be one that, at the same time that recovers information by noise removal, also preserves shapes and edges, that are very important for texture analysis. Given that, the search for the filter to be used was done in the field of Nonlinear Filters, which are filters whose output is based not only in one pixel but in the image region that is within the filter window.

Image filtering was done through a median filter, using a 3x3 neighborhood with zero-padding - to account for neighborhoods that fell outside the image. Median Filtering is usually used in the task of random noise reduction, as it is the case of this work, while preserving image's edges, as it is needed [75]. When applying median filters to an image, a window of previously defined dimensions (3x3, in the present case) slides across the image, and the intensity of the pixel being processed becomes the median value of the pixels inside the window.

An example of a median filter application in one pixel is depicted in Figure 3.6.

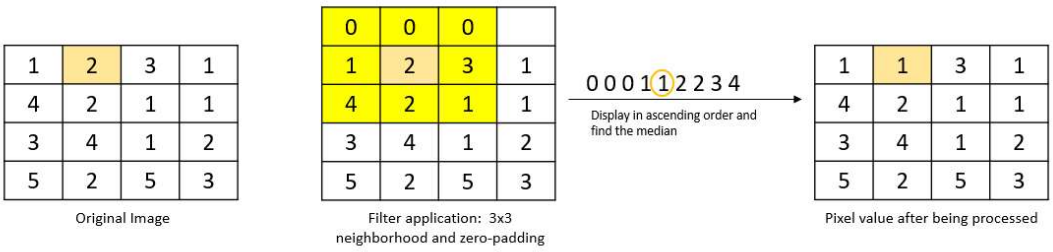

Figure 3.8- Example of Median Filtering application with a 3x3 neighborhood, and zero-padding. The resulting filtered image is shown on the right.

Noise reduction effect is similar to what happens in low-pass filtering; however, median filtering has two great advantages: discontinuities can be preserved, and pixels that are significantly different from their neighborhood can be smoothen without disrupting surrounding pixels. These two advantages allow that while image noise is reduced, its edges and shapes are, until a certain limit, preserved. For the purpose of this work, which is based on textural analysis, it is expected that noise addition disrupts the image to a point where no textural characteristics can be assessed to differentiate cancer and non-cancer cases. Nonetheless, that was the reason that served as motivation to pursue a

median filtering approach. The application of a filtering technique has the goal of, with noise reduction and shape preservation, recover textural information that was lost with noise adding, allowing a fair classification. The median filter was applied to the noisy images across all noise magnitudes. It is important to note that a copy of the noisy images before filtering was maintained in order to be used in further studies.

Across this work, when assessing images that have noise and are not filtered, they will be referred to as noisy/noised images. On the other hand, when dealing with the images that were filtered after noise adding, they will be referred as filtered images.

To check for a correct procedure, three different noise metrics that compared noisy images/filtered images with the original images were computed. These were: Mean Squared Error, peak Signal-to-Noise Ratio, and Structural Similarity.

As the name implies, Mean Squared Error (MSE) is, in this scenario, a measure of the averaged squared difference between the original image and the noised image/filtered image. MSE is commonly used because the error between the two images might be negative [76] and, by squaring the difference between each pair of images, that problem is avoided. Besides that, MSE is of very simple calculation, requiring low computational cost, and has a very clear physical meaning, which makes it a desirable image quality metric. Since it is an error measure, a lower value of MSE represents a higher similarity between images.

Considering I_n as the noised image , I as the original image, and n as the number of elements in the image, MSE was computed as shown in equation 3.3.

$$MSE = \frac{\sum_i (I_n(i) - I(i))^2}{n} \quad (3.3)$$

Peak Signal-to-Noise Ratio (pSNR) has a computation slightly different to the usual SNR. This metric aims to express how much noise affects image representation, by performing the ratio between the maximum signal value and the maximum noise power. As it happens to MSE, there is also a clear physical meaning associated with pSNR, and the computation is also simple. Nonetheless, this is only a powerful tool if the images that are being compared have different dynamic ranges, which is not the case, since image normalization is being performed. For that reason, this measure is included in this work only to validate what is presented by the MSE metric and will not be deeply explored.

The formula presented in equation 3.4 was used to compute pSNR [77] :

$$pSNR = 10log_{10}\frac{(peak\ value)^2}{MSE} \quad (3.4)$$

Finally, Structural Similarity (SS), which some authors point out to be an improvement to MSE and to pSNR, measures image quality based in three [78] different components: luminance, contrast and structure. This is a perception-based model, inspired in the human visual system. Each of these three components was computed as shown in equations 3.5-3.7.

$$luminance\ (x,y) = \frac{2\mu_x\mu_y + 10^{-4}}{{\mu_x}^2 + {\mu_y}^2 + 1\ \ ^{-4}} \quad (3.5)$$

$$contrast\ (x,y) = \frac{2\sigma_x\sigma_y + 9\times10^{-4}}{{\mu_x}^2 + {\mu_y}^2 + 9\times10^{-4}} \quad (3.6)$$

$$structural(x,y) = \frac{\sigma_{xy} + 4.5\times10^{-4}}{\sigma_x\sigma_y + 4.5\times10^{-4}} \quad (3.7)$$

With μ_x, μ_y being images' means, σ_x, σ_y being images' standard deviation, and $\sigma_x\sigma_y$ being images' cross-covariance. Lastly, SS measure was calculated using equation 3.8.

$$SSIM(x,y) = luminance(x,y) \cdot contrast(x,y) \cdot structural(x,y) \quad (3.8)$$

Going further in pre-processing steps, most of the works done in texture analysis of mammograms define a small ROI to assess texture. Despite that, as it was mentioned in the previous section, some authors advocate that using a single ROI does not account for texture heterogeneity across the entire breast. Although some authors tried to overcome this problem by considering multiple ROIs across breast tissue – with positive results - none of the researched papers considered the approach used in this thesis: recurring to the entire breast, as a whole, to assess textural patterns.

When looking to the images in the dataset used in this study, it became clear that images from *Hospital da Luz* and from the public dataset had very different sizes, and it was important to assess if breast and/or image size could, in some way, bias the classifier performance. In fact, what was really important was to evaluate not only the absolute image size, but also the breast tissue size – it's % of occupation in the analyzed images. In order to do that, after image cropping was performed, image binarization was done and the number of "on" pixels was counted. To calculate the percentage of occupation, a simple calculation was done: the number of "on" pixels was divided by the total amount of pixels in the image. Having these measure, after classifier construction and testing, it was possible to verify if breast dimension were, in any way, affecting classifier's performance.

3.3 - Feature extraction

After granting a correct data pre-processing, images are ready to be processed in what concerns to texture evaluation. The next step towards classification is feature extraction. Breast parenchyma characteristics that relate to texture were extracted, analyzed, and used as features to be fed to classifiers. Absolute grey level, Haralick and Run-Length features were mathematical retrieved from each mammogram. Once again, contrarily to what happens in other research, image analysis will be performed in the entire breast region and not in a single ROI. This methodology allows a wider view of the entire breast tissue, in opposite to what happens in a single ROI analysis, where only a small breast segment is being analyzed. As presented in the introduction section, there are several feature groups that can be computed and correctly describe the parenchyma. Considered the scope of this work, only the most used features were be computed. For feature calculation, both self-written functions and MATLAB toolboxes were used.

In what concerns to absolute grey-levels/histogram features, four features were computed: the first two (Average and Variance) related to the absolute grey-levels of each image, and the following two features (Skewness and Kurtosis) computed through a defined intensity histogram. A brief description and the formula used to compute the respective feature can be found in Table 3.1.

Feature	Description	Formula ($x_i \rightarrow i^{th}\ pixel\ value, N \rightarrow n^{o}\ of\ pixels$)
1. Average	Expected pixel value	$f1 = \frac{\sum_{i=1}^{N} x_i}{N}$
2. Variance	Expected value of the squared deviation from the mean	$f2 = \frac{\sum_{i=1}^{N}(x_i - f1)^2}{N}$
3. Skewness	Measure of asymmetry (probability distribution) around the mean	$f3 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - f1)^3}{(f2)^{\frac{2}{3}}}$
4. Kurtosis	Measure of flatness of the probability distribution.	$f4 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - f1)^4}{(f2)^2}$

Table 3.1 - Histogram Features Table: with feature name, brief description and formula used.

For Haralick texture analysis, features are not directly extracted from pixel intensity values in the mammogram. Before starting with the actual feature extraction, normalized images were used to compute the gray-level-cooccurrence matrix (GLCM). Four different matrices were calculated, one for each direction that was used to search for intensity pairs – 0º, 45º, 90º and 135º. Therefore, each feature have four variations, one for each matrix. As it is common practice to quantize grey-levels when computing the GLCM, in this work it was defined that for every matrix construction the normalized image was quantized to 256 grey-levels.

In the further feature description, $p(i,j)$ represents the (i,j) entry of the normalized GLCM and, opposite to what happens in absolute gray-levels calculations, now, *N* represents the number of existent gray-levels. Table 3.2 addresses textural features retrieved from the GLCM. Although there are eighteen Haralick features, only eight will be accompanied by a brief description once the remaining are derived from these ones.

Besides that, other notations should be addressed, for feature computation:

1. $p_{x+y}(k) = \sum_{i=1}^{N}\sum_{j=1}^{N} p(i,j)$, with $k = i + j$, ranging from 2 to *2N.*
2. $p_{x-y}(k) = \sum_{i=1}^{N}\sum_{j=1}^{N} p(i,j)$, with $k = |i - j|$, ranging from 0 to *N-1.*
3. $p_x(i) = \sum_{j=1}^{N} P(i,j)$, with P being the non-normalized GLCM.
4. $p_y(j) = \sum_{i=1}^{N} p(i,j)$.
5. $\mathrm{HXY} = -\sum_{i,j} p(i,j) \log(p(i,j))$, entropy of *p(i, j).*
6. HX, HY, entropy of p_x and p_y.
7. $HXY1 = -\sum_i \sum_j p(i,j) \log(p_x(i)p_y(j))$.
8. $HXY1 = -\sum_i \sum_j p_x(i)p_y(j) \log(p_x(i)p_y(j))$.
9. $Q = \sum_k \frac{p(i,k)p(j,k)}{p_x(i)p_y(k)}$.

Feature	Description	Formula ($x_i \rightarrow i^{th}\ pixel\ value$)
5. Energy (or Angular Second Momentum)	Measure of Homogeneity/Uniformity	$f5 = \sum_{i,j} p(i,j)^2$
6. Correlation	Measure of correlation between each pixel and its neighbor	$f6 = \frac{(i - f1i)(j - f1j)p(i,j)}{\sigma_i \sigma_j}$
7. Contrast	Measure of local pixel intensity variation	$f7 = \lvert i - j \rvert^2\ p(i,j)$
8. Inverse Different Moment	Measure of local Homogeneity	$f8 = \sum_{i,j} \frac{p(i,j)}{1 + (i-j)^2}$
9. Sum of Squares	Measure of how different an element of the matrix from the mean is.	$f9 = \sum_{i,j} (i - \mu)^2\ p(i,j)$
10. Entropy	Measure of Randomness.	$f10 = -\sum_{i,j} p(i,j) \log\ (p(i,j))$
11. Sum Average	-	$f11 = \sum_{i=2}^{2N} i\ p_{x+y}(i)$
12. Sum Entropy	-	$f12 = -\sum_{i=2}^{2N} p_{x+y}(i) \log\ (p_{x+y}(i))$
13. Sum Variance	-	$f13 = \sum_{i=2}^{2N} (i - f12)^2\ \ p_{x+y}(i)$
14. Difference Variance	-	$f14 = \mathrm{VAR}(p_{x-y})$
15. Difference Entropy	-	$f15 = -\sum_{i=0}^{N-1} p_{x-y}(i) \log\ (p_{x-y}(i))$
16. Information Measure of Correlation I	-	$f16 = \frac{HXY - HXY1}{\max\ (HX, HY)}$
17. Information Measure of Correlation II	-	$f17 = (1 - \exp[-2(HXY2 - HXY)])^{\frac{1}{2}}$
18. Maximal Correlation Coefficient	-	$f18 = (2^{nd}\ largest\ eigenvalue\ Q)^{\frac{1}{2}}$
19. Cluster Shade	Measures of Asymmetry. High values indicate that the image is non-symmetric.	$f19 = \sum_i \sum_j \left(i + j - \mu_i - \mu_j\right)^3 p(i,j)$
20. Cluster Prominence		$f20 = \sum_i \sum_j \left(i + j - \mu_i - \mu_j\right)^4 p(i,j)$
21. Dissimilarity	-	$f21 = \sum_i \sum_j \lvert i - j \rvert\ p(i,j)$
22. Max probability	-	$f22 = \max\ (p(i,j))$

Table 3.2- GLCM features Table: with features name, brief description and formula used.

Finally, in what concerns to Run-Length features, as it happens to GLCM features, extraction did not occur straight from each image, but rather from RLM. Here, this matrix was also constructed considering the normalized images, that were quantized to 16 gray-levels. Each feature should have a

different value depending on the direction of construction of the RLM, however, an approach that combined the information across all directions (0º,45º,90º,135º) was used. Therefore, each feature had only one value. A brief description and the formulas use to compute each feature, independently of the search direction, are present in Table 3.3 and the following notations should be considered when studying the formulas:

1. $p(i,j)$, it's the (i, j) entry of the Run-Length Matrix.
2. N_g, is the number of gray-levels in the original image.
3. N_r, number of different occurring run-lengths.
4. P, number of pixels in the image.

Feature	Description	Formula ($x_i \rightarrow i^{th}\ pixel\ value$)
23.Short Run Emphasis	Metric that grows when the image has more short runs	$f23 = \frac{\sum_i^{Ng}\sum_j^{Nr}\frac{p(i,j)}{j^2}}{\sum_i^{Ng}\sum_j^{Nr}p(i,j)}$
24. Long Run Emphasis	Metric that grows when the image has more long runs	$f24 = \frac{\sum_i^{Ng}\sum_j^{Nr}j^2 p(i,j)}{\sum_i^{Ng}\sum_j^{Nr}p(i,j)}$
25. Gray-Level Nonuniformity	Metric that grows when gray-level outliers dominate the image	$f25 = \frac{\sum_i^{Ng}(\sum_j^{Nr}p(i,j))^2}{\sum_i^{Ng}\sum_j^{Nr}p(i,j)}$
26. Run Percentage	Measure of image homogeneity.	$f26 = \frac{\sum_i^{Ng}\sum_j^{Nr}p(i,j)}{P}$
27. Low Gray Level Run Emphasis	Measure that increases when the image as more runs of low gray-levels	$f27 = \frac{\sum_i^{Ng}\sum_j^{Nr}\frac{p(i,j)}{i^2}}{\sum_i^{Ng}\sum_j^{Nr}p(i,j)}$
28. High Gray-Level Run Emphasis	Measure that increases when the image as more runs of high gray-levels	$f28 = \frac{\sum_i^{Ng}\sum_j^{Nr}i^2 p(i,j)}{\sum_i^{Ng}\sum_j^{Nr}p(i,j)}$

Table 3.3- Run-Length Features Table: with feature name, brief description and formula used.

After computing the entire set of features for each case of the dataset, another matrix was constructed, where each line represented a case of the dataset and each column represented one of the features that was retrieved. After that, and considering the evaluation present in the dataset, an additional column was added to the matrix, representing the labels – '1' for cancer cases and '0' for normal cases.

The process discussed in this subsection was repeated for original, noised, and filtered images.

3.4 - Feature Selection

Feature selection methodologies are important once, by removing non-relevant/redundant features from the constructed set, problem's dimensionality is reduced, which leads to low computational power needed, and a more efficient classifier performance.

The first task immediately before feature selection is dataset construction, using the extracted features, into train and test matrices. This was done in a way that 70% of the cases were used for training and 30% were left out to be used in the testing phase – in this scenario, as explained, four different test sets were created, maintaining the 70/30% ratio.

In an early stage of this work, feature selection was aimed using common variable selection methodologies, like stepwise feature selection, using built-in MATLAB functions. However, this procedure required not only high computational power, as also revealed itself as inefficient, with only

one to two features being selected as significant and, consequently, resulting in classifiers with very poor performance. Given that, a shift in rationale was made and, instead of looking for automatic feature selection, a dive into feature scoring was made, and two possible alternatives appeared: 1) Using, once again, built-in MATLAB functions to score features according to some statistical criteria or 2) using a software called Orange [79]. This is an open-source ML and data mining tool, that allows users to perform visual programming and interactive data visualization. In what concerns to feature selection, the software provides a widget called 'Rank' to perform it. Having in mind the time-consumption and inefficiency of MATLAB software when using sequential feature selection, and considering the visual simplicity presented by Orange, this software was chosen to be used in feature selection phase. Orange allows users to import csv files consisting of columns of features, with each row representing a case from the dataset and with the last column representing labels associated to each case – these files represent the matrices explained in the beginning of this subsection. Once the file is imported, it can be directly fed to the Rank widget, which will output a ranking of features, with a score associated to each predictor. The user can, using this widget, choose which features to consider. The presented scores represent how much each variable is correlated with the label and can be computed based on different criteria. For the purpose of this work, three criteria – that are filtering methodologies – were considered:

1. Relief-F: Measures how well a feature can discriminate similar cases with different classes.
2. Chi: Measures how much each feature and the target class are related, through a chi-squared test.
3. Information Gain: As the name implies, represents the amount of information that is gained with each feature.

The process of feature selection included not only the *rank* widget but also *test and score* and *confusion matrix* widgets. The first widget will receive as input the data form the csv file, but only with the columns that represent the chosen features, and it will train and test embedded classifiers. As for the second widget, it will receive as input the trained classifiers and output the respective confusion matrices.

Feature selection methodology was done three times, one for each criterion, and tracked the following steps:

1. Use the entire feature set to train and test the embedded classifiers.
2. Retrieve AUC values for each classifier, as well as confusion matrices.
3. Disregard features that had less than one-third of the maximum score.
4. Use the subset of selected features to train and test the embedded classifiers.
5. Repeat point 2.
6. Compare AUC and k-coefficient values between the entire feature set and the relevant feature set.

K-coefficient is a different statistical measure that is directly derived from the confusion matrix of each classifier. This coefficient allows to understand the "strength of agreement" [80] between predicted and real classes. First, it is important to understand how this measure is calculated and, for that purpose, consider the following matrix (Table 3.4):

		Predicted Classes	
		0	1
Real Classes	0	10	2
	1	3	9

Table 3.4 - Confusion Matrix Example

Given this matrix, k-coefficient is computed using equation 3.9.

$$k = \frac{p_o - p_e}{1 - p_e} \quad (3.9)$$

With p_o and p_e being:

$$p_o = \frac{correct\ 0's + correct\ 1's}{all\ cases} = \frac{19}{24}$$

$$p_e = \frac{real\ 1's \times predicted\ 1's}{(all\ cases)^2} + \frac{real\ 0's \times predicted\ 0's}{(all\ cases)^2} = \frac{1}{2}$$

This results in:

$$k \approx 0.58$$

The question resides in how to interpret this *k* value, and that is where the work of Landis and Koch enters. These authors defined classes of agreement for specific ranges of *k* values. Those classes can be defined as seen in Table 3.5.

k-coefficient	Strength of Agreement
< 0.00	Poor
0.00 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost Perfect

Table 3.5 - Strength of Agreement classes. Adapted from [80]

After understanding this metric, feature selection was finished by comparing values presented by *AUC* and *k-coefficient* before and after feature removal. If, after the feature removal done in point 3, *AUC* was maintained or increased, and if *k-coefficient* stayed in the same class or improved, the removed predictors were considered non-significant and, therefore, were not taken into account for further phases of the project, being removed from the training and testing sets. This was the used method because in feature selection what is expected is that, when non-relevant features are eliminated from the considered feature set, classification performance is not significantly altered in a negative direction. Thus, when comparing AUC and *k* coefficient before and after feature selection, the values should remain close to the original, or even increase.

The results obtained by different criteria will also be compared and analyzed, in order to choose the best feature set.

The use of three criteria is done to improve the probability of choosing a better feature set. By using only one criterion, the only comparison to be made would be between the before and after feature selection, but with this methodology, a comparison between different selected feature sets can also be made. The outline of the orange toolkit used is present in Figure 3.7.

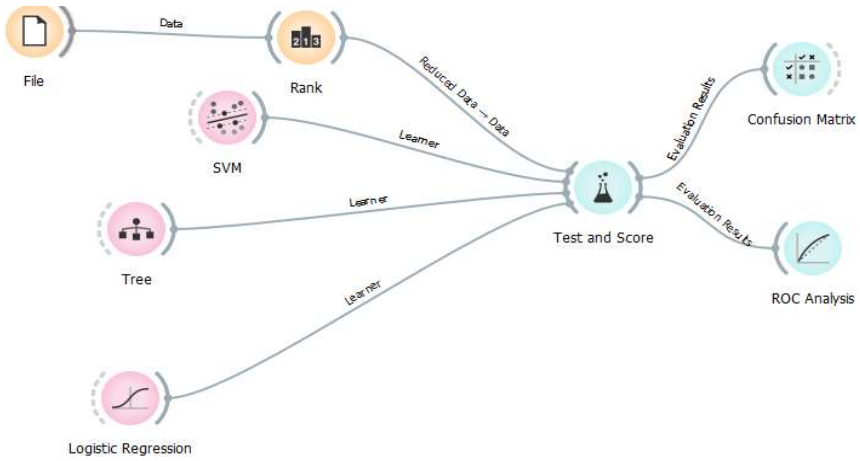

Figure 3.9- Orange Software outline used in this thesis.

3.5 – Classification Phase

After the optimal set of features was chosen, features extracted from images in the training set were given as input to a ML algorithm, along with labels: '1' for cancer cases and '0' for healthy cases. There are several classifiers that can be used to perform the task of differentiating between cancer and healthy patients and, in order to avoid bias by choosing a specific classifier *a priori*, four different algorithms were used: Support Vector Machine (SVM), Decision Tree (DT), Logistic Regression (LR) and Discriminant Analysis (DA). A brief explanation of each of these classifiers is presented below.

3.5.1 - Support Vector Machines (SVM)

SVM is an algorithm widely used for solving both classification and regression problems and it is specially formulated for binary classification [81]. This is an algorithm that aims to find the hyperplane that better separates the data cases from each class. The best hyperplane is the one that maximizes the margin between the two classes, i.e., the one that maximizes the distance between the hyperplane and the closest points to the hyperplane from each class. These closest points are called support vectors and are very important to define the location of the hyperplane and consequently maximize the margin. When the cases are not perfectly separable, this algorithm will map data to a higher dimension, where data separation is no longer a problem. This mapping to a higher dimension, in order to find a hyperplane that can separate the data, is done using kernel functions that can be one of three types: Gaussian, Linear and Polynomial. Considering x_i and x_j two feature vectors in feature space *X*, kernel formulas can be defined as shown in equations 3.10-3.12.

$$\text{Gaussian: } K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}} \quad (3.10)$$

$$\text{Linear: } K(x_i, x_j) = x_i{}^T x_j \quad (3.11)$$

Polynomial: $K(x_i, x_j) = (x_i{}^T x_j + 1)^p$ *(3.12)*

In this thesis, for SVM, four different hyperparameters were optimized: Box Constraint, Kernel Scale, Kernel Function and Polynomial Order. The first parameter is related to the idea that the data is not perfectly separable, and, for that reason, the algorithm should allow some mis-classifications in the process of finding the best hyperplane. The higher the value of this parameter, the higher the cost of misclassification. So, a higher value leads to a more strict separation between classes. Kernel Scale is a scaling parameter applied to the input before kernel application. The default value is one but can assume any positive scalar value. All values of the input (feature) matrix are divided by the Kernel Scale before the kernel is applied. Kernel Function is, as the name implies, the type of kernel that is chosen for data mapping to higher dimensions and it can be either gaussian, linear or polynomial. If a polynomial kernel is chosen, then a fourth hyperparameter needs to be chosen, and that is the order of the polynomial that it is going to be used.

3.5.2 - Decision Trees (DT)

Decision Tree algorithm can be understood as a flowchart with different branches, where data is going to be continuously separated until each dataset case is allocated to a specific class. This is called a decision tree because it is composed by different nodes, where the first one, that has no incoming edges, is called root node. Data splitting at each node will be done based on feature values, either numerical or categorical. The nodes in the middle of the tree, that have both incoming and outcoming edges are called internal nodes. Finally, the last nodes of the tree, that have no outcoming edges, are called terminal nodes (or leaves) and it is where the decision of allocating each case to a class is made. The construction of this classifier starts by partitioning the entire data based in the feature that provides lowest impurity values. The process of consecutively splitting the data, according to a specific splitting criterion, will be done until all elements of the same leaf node have the same classification [82]. When having new data instances, they will enter the root node, pass through internal nodes based in their feature values and arrive a leaf node that will classify them.

MATLAB allows two different splitting criteria, based on impurity measures, that are called: Gini Diversity Index and Deviance, where pure nodes have a value of zero for each of these metrics. Calculations for these measures can be performed as shown in equations 3.13 and 3.14.

$$Gini: 1 - \sum_i p^2(i) \quad (3.13)$$

$$Deviance: -\sum_i p(i) log_2\, p(i) \quad (3.14)$$

The sums present in these formulas are done across the classes and *p(i)* is the fraction of cases with class *i* that reach the node. A general depiction of a decision tree architecture can be seen in Figure 3.8.

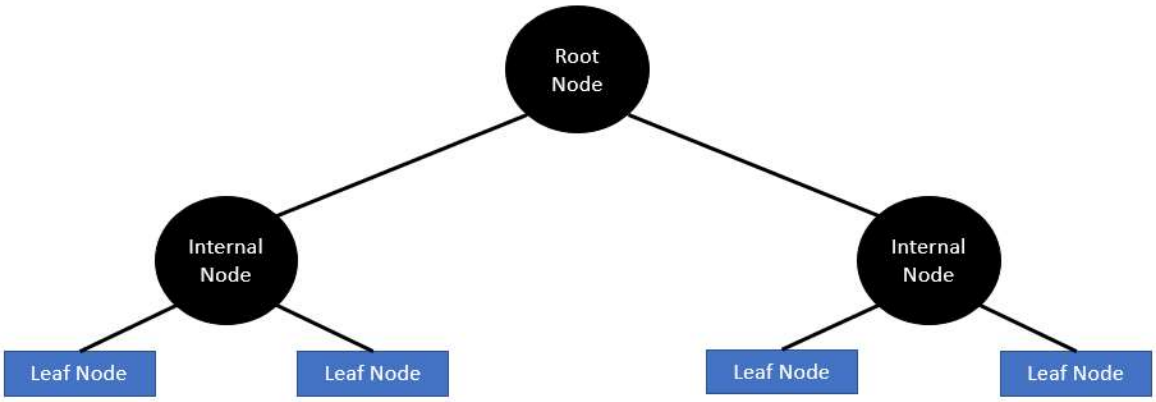

Figure 3.10 - Decision Tree Architecture

For *Decision Tree,* three different parameters were optimized: Maximum Number of Splits, Minimum Leaf Size, and Split Criterion. Maximum Number of Splits represent the maximum possible number of decision splits that can come out from an internal node when constructing the architecture of the tree. On the other hand, Minimum Leaf Size represents, as the name implies, the minimum number of observations that are allowed to be present in each leaf node. Lastly, splitting criteria refers to the selection of what criterion to use when choosing what is the best possible splitting to do in the development of the Decision Tree algorithm. Either Gini index or Deviance can be chosen in this hyperparameter.

3.5.3 - Logistic Regression (LR)

Logistic Regression is a commonly known statistical model that relates the probability of an event with outcomes '0' or '1' with a previously defined set of predictors. These predictors are, in the case of this work, the selected feature set and have associated coefficients that need to be estimated. Usually, maximum likelihood estimation is performed in order to correctly describe the relation between the predictors and the outcome of the model. Logistic regression will then, based in the values that each feature assumes, compute the probability of each case to belong to the class '0' or '1' and, using a probability threshold, a classification is made. Assuming *p1* as the probability of a case to belong to the '1' class; β's as the model parameters; and *X's* as the model predictors, the logistic regression model can be described as follows – equation 3.15:

$$\ln\left(\frac{p1}{1-p1}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n \quad (3.15)$$

Authors point out that the use of Logistic Regression in this type of applications is better than the use of Linear Regression, once less requirements and assumptions are imposed by Logistic Regression. For example, Logistic Regression does not assume that the relation between the outcome ('0' or '1') and the predictors is linear and does not require normally distributed features [83].

3.5.4 - Discriminant Analysis (DA)

DA is one of the most used algorithms in classification, firstly introduced by R. Fisher in 1936. This method uses discriminant functions, one for each class, to determine decision boundaries. A brief outline of the algorithm is given by [84] and it goes as follows: given a feature matrix as input (with the respective labels), the mean of each class is calculated; after that, the prior probability and the covariance matrix for each class is also computed. Once these steps are defined, and recurring to

their results, discriminant functions are calculated, and decision boundaries are defined.

In order to classify an unknown case, the value of that case should be assigned into the discriminant functions of each class. The class given to the case is going to be the one with the highest discriminant function value.

There are two types of DA Classifiers: the first and most simple occurs when an assumption is made that all classes have the same covariance matrix, which results in linear decision boundaries, hence being called Linear Discriminant Analysis; when each class has an independent covariance matrix, the decision boundaries becomes quadratic, hence resulting in a Quadratic Discriminant Analysis [84].

Finally, for DA, two hyperparameter were optimized: Delta, and Discriminant Type. Delta is a linear coefficient threshold which means that if any coefficient (value associated to feature) in the model has a value smaller than delta, this coefficient becomes zero, eliminating the predictor that is associated with it. The higher the delta value, the higher is the amount of eliminated predictors. Discriminant Type can vary between six options that are related to the Covariance matrix of each class. As it was already perceived, when the covariance matrices vary among classes, Discriminant Type assumes the value 'quadratic'. When this covariance matrices are diagonal and can vary among classes, the hyperparameter becomes 'diagquadratic'; if the matrices can vary between classes and the software inverts the matrix using a pseudo inverse, the value of the parameter is 'pseudoquadratic'. The same rationale can be applied for the case where all classes have the same covariance matrix but here the hyperparameter assumes as values: 'linear', 'diaglinear', or 'pseudolinear'.

3.5.6 – Classifier Optimization Options

The first step towards classification was training each of the explained classifiers, recurring to the training dataset. During the training phase, and in order to achieve the best possible performance, hyperparameter optimization routines were implemented for each classifier. To choose the best parameter configuration set, as referred in section 1.8, cross-validation performance measures were considered. Each classifier, except for Logistic Regression which was not optimized due to MATLAB constrictions, has different parameters that need to pass through optimization. A brief description of each optimizable parameter considered in this study was addressed in the previous subsections.

Nonetheless, the use of optimization routines has also some parameters that need to be selected in order to perform the said optimization. Since different optimization parameters result in different classifier performance, and to assure that the effects of fixed *a priori* choices are minimized, 10 variations for each classifier, with different optimization parameters, were trained, validated, and tested. The first parameter is the Optimizer function, that can vary amongst Bayesian Optimization, Grid Search and Random Search. When the chosen optimizer is Bayesian, then an Acquisition Function should also be defined. Another optimization option is Max Objective Evaluations, i.e., the number of iterations in the optimization process. This value was left to default (30) for Bayesian and Random search and was switched between 100 and 500 for Grid Search. This choice was done based on a preliminary classification task, where it was found that the value 30 provided positive results for Bayesian and Random search, that were not improved when this value was increased. Grid Search with a value of 30 in this parameter performed poorly and good results were only seen when the Max Objective Evaluations parameter was highly increased. For each classifier's variation, the evaluation of the chosen hyperparameter set was done twice, one with 5-fold cross-validation and another with 10-fold cross-validation. An outline of the different classifiers' – except for LR - variations is

presented in Table 3.6.

	Optimizer	Acquisition Function	Max. Iteration	Validation
Variation 1	Bayesian	Expect. Improv/sec	-	10-fold
Variation 2	Bayesian	Expect. Improv/sec	-	5-fold
Variation 3	Bayesian	Expect. Improv.	-	10-fold
Variation 4	Bayesian	Expect. Improv.	-	5-fold
Variation 5	Grid-Search	-	100	10-fold
Variation 6	Grid-Search	-	500	5-fold
Variation 7	Grid-Search	-	100	10-fold
Variation 8	Grid-Search	-	500	5-fold
Variation 9	Random-Search	-	-	10-fold
Variation 10	Random-Search	-	-	5-fold

Table 3.6 – Classifiers' Variations in terms of Optimization Options- Optimizer used, Acquisition Function defined when the optimizer is Bayesian, Maximum Allowed Optimization Iterations, and k Cross-validation folds.

It is, therefore, important to have a basic understanding of what each Optimizer function represents and what varies between them.

Starting with Grid Search, this is an optimization algorithm also called parameter sweep, once it is a method of exhaustively searching for the best parameter setting possible amongst a pre-defined set of parameter values. For example, for SVMs there are three to four possible optimizable hyperparameters, so, for this optimization algorithm, a number of possible values for each hyperparameter is previously defined (10, by default) and all different combinations of these parameter values are tested until the best one is found. It can be noted that this is a computationally heavy algorithm, and very time consuming, which explains the limitation of the "Max Objective Evaluation" parameter.

Random search is similar to grid-search but instead of testing all possible combinations amongst hyperparameters it randomly chooses combinations until a stopping criterion (Max Objective Evaluation) is reached. As it is perceived, Random Search is far less heavy than Grid Search, with the downside that there could be an optimal hyperparameter set that was missed.

Bayesian Optimization is one of the most used optimization routines for unknown expensive functions. This is an iterative algorithm that, in this case, was repeated thirty times, based on two principles: The use of a probabilistic (gaussian) surrogate model, that models the unknown function, and an acquisition function that is going to be used to determine which point to analyze in the following iteration. The goal is to fit the surrogate model into the observations of the target function that were already made and after that, the acquisition function, through the probability distribution of the model, will find the next point to be evaluated, balancing the decision of evaluating new points in unexplored areas or evaluating points in promising areas that were found in previous exploration. There are several options for acquisition function but, in order to reduce the dimensionality of the problem, only the ones that provided better results in the preliminary testing task were used: Expected Improvement and Expected Improvement Plus[85]. Each classifier variation was trained and tested with the previously defined and referred sets. The AUC was the metric used to evaluate each classifier variation. The classifier that presented a better performance was chosen and all the studies related to classification were only performed in a specific optimized classifier.

3.5.7 – Further Classifier Studies

After choosing the classifier, studies related to density measures and area calculations were performed. The goal in this stage is to verify if either density or breast size bias classifier decisions when allocating each case to a class label. While in terms of density what was aimed was to check how the classifier performs across different BI-RADS classes; for the case of area influence, the aim was to check if the class given by the classifier and the image/breast area were, somehow, correlated.

To assess area influence, the test set that resulted in the best classifier performance, amongst the four classifiers implemented, was studied. First, a visual search for patterns in area distribution across classes and classification was performed. After that, correlation coefficients between image area and output class, as also between breast % area and output class were computed.

For density studies, classifier AUC for each BI-RADS class was evaluated, in order to assess if there are some specific densities that make the classifier to have a poor performance. In order to do that, different test sets, each one containing images from just one density class needed to be created. However, two problems arise: 1) Luz data did not have any information about density and could not be considered; 2) there was only one healthy image with BI-RADS 4 class. For these reasons, classifier performance was only evaluated for classes 1,2, and 3, and testing set construction was done as follows.

1. Considering the available images for each BI-RADS class, the number of images to be considered for each test set were defined: N1=15 images; N2=N3=35 images.
2. Using the available CBIS-DDSM cases, 3 feature matrices were created – each matrix containing cases of one of the BI-RADS classes.
3. 15 and 35 random rows were respectively randomly drawn from each of the matrices, resulting in three test sets, one per BI-RADS class.
4. AUC was computed for each of the three test sets.
5. To avoid bias, points 3 and 4 were repeated five times – resulting in five different test sets for each density class.

Besides using textural features, some articles analyzed in subsection 2.5.4, computed, and used %PD as a feature. Breast density presents itself as an important risk factor, and %PD is a mammography-based calculation that is directly related to breast density. Therefore, the goal here was to repeat the training and testing procedure of classifiers but considering the selected texture features along with %PD. Then, it would be possible to verify if this new classifier, that includes information about breast density, outperforms the previously considered classifier.

%PD calculation was done through an application developed by Madalena Simões in her Master Thesis [86], and the process by which this was done will be briefly described further. To compute the percentage of the breast that is dense, one must first achieve a differentiation between dense and non-dense tissue, or in other words, fat tissue and fibroglandular tissue. Then, %PD is simply the ratio between fibroglandular and adipose tissue. The differentiation between these tissues is done through segmentation, that usually uses techniques like k-means clustering algorithm. In general, the algorithm will allocate each pixel to a cluster – fatty or fibroglandular – based in its intensity value, meaning that similar pixels get to be in the same cluster. Pixel clustering is done in a way that allows intra-cluster variance to be minimum, while inter-cluster variance is maximum. A variation of this algorithm, called fuzzy c-means, is also widely used, with the difference that each pixel can belong to more than one cluster, having a degree of membership to each of the clusters. This degree is usually a distance metric from the analyzed pixel to the center of the cluster and, the lower the distance to a cluster, the higher the membership. This was the technique used by Madalena to divide the pixels into a fatty and fibroglandular cluster, in order to compute %PD.

Finally, using the chosen classifier an Interactive Application that allows the user to load an image, pre-process it and classify it, was developed.

3.6 - Risk Assessment

In order to assess BC risk using ML techniques, in a fashion like the ones explored in section 2, one needs sequential year mammograms and/or information about future development of BC. However, the method by which the dataset is divided can vary widely.

Initially, considering the use of ML algorithms, there were two possible lines of work to pursue:

1. Consider *each image* in the dataset as a single case and classify as high-risk the ones that were taken immediately before a cancer diagnosis.
2. Consider *each woman* as a case, and, if there had been a cancer diagnosis, all the images of that women were considered for the high-risk group.

It should be noted that, for scenario 2, the features from each image would not be used directly to train classifiers, but rather variations of them that represented their fluctuations across different years. Scenario 1 had shown to provide positive results in the literature [70], being simpler, and using methodologies already applied for cancer/non-cancer differentiation. However, the idea that the use of mammograms across years might help to discover useful information led to the formulation of scenario 2.

Only Hospital da Luz data has information about further cancer diagnosis and, for that reason, only those images could be used. The problem is that there are only 4 patients that ultimately developed BC, in a total of 11 patients. The small dimension of the dataset did not allow for the use of any ML technique. However, and since risk analysis showed to be a very promising area, a visual analysis of possible patterns in feature values across years for both high-risk and low-risk groups could be a line of work to consider.

Since direct feature evaluation from each image would be extremely difficult, feature value fluctuation analysis across years was aimed. Besides that, and to help the visual analysis of these fluctuations, the following six metrics were computed using the originally extracted features from every image.

1. Absolute difference between first and last year.
2. Absolute difference between first and last year, divided by the number of years that differ from first to last mammogram.
3. Absolute difference between last and immediately previous year.
4. Absolute difference between last year and two years before.
5. Mean of the absolute difference between sequential years.
6. Mean of the absolute difference between sequential years, divided by the number of years that differ from first to last mammogram.

Apart from computing absolute differences, the last six metrics were calculated using structural similarity, through the available images.

Although feature differences calculations could occur without problem, the same cannot be said about structural similarity calculations. In order for structural similarity to have a meaning when comparing two imagens, it is important that the pixels being compared are representative of roughly the same structures. In order to achieve that, image registration needed to be conducted. This process,

that it is concern with image alignment, was done automatically in an intensity-based fashion, which is an iterative process that uses: two images, a fixed reference image and a moving image; a metric that describes how similar the images are; a registration optimizer; and a transformation type, that will align the moving and the fixed images.

The process by which the registration occurs is displayed in Figure 3.9 and it can be summarized as follows:

1. Based on a chosen transformation type, a matrix transformation is defined.
2. The transformation is applied to the moving image.
3. A metric that compares the transformed moving image and the fixed image is computed.
4. A chosen optimizer verifies if a stopping criterion (a decrease in metric value in relation to previous iterations, a pre-defined iteration limit, etc.) has been met. If it is met, then the process is finalized, and the moving image is said to be registered. In the negative case, the optimizer rearranges the transformation matrix, and the process is repeated.

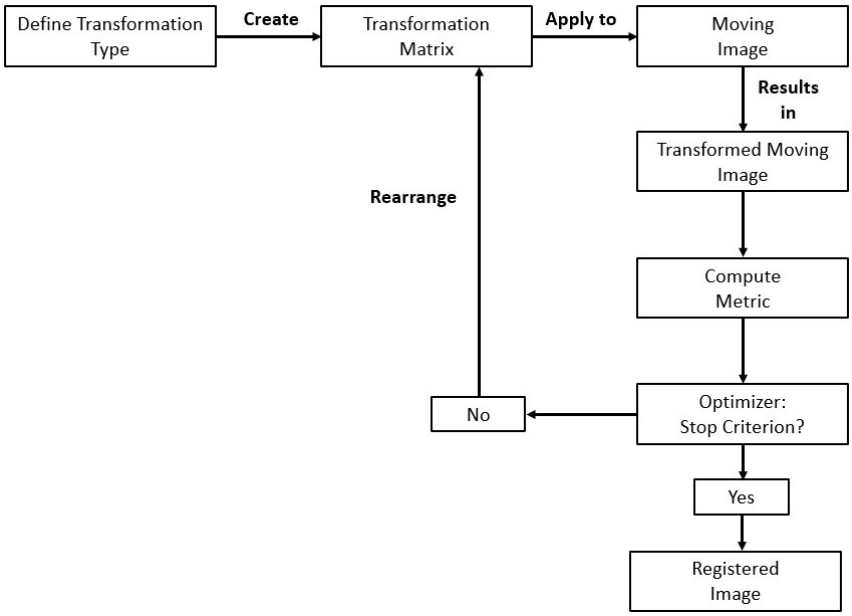

Figure 3.11 - Image Registration Methodology Summary.

The process of choosing the metric and the optimizer can be difficult and an automated method was implemented in MATLAB. This method asks as input if the images that are going to be registered are *Monomodal* or *Multimodal.* All images in the dataset are mammograms, what should point that the input should be *monomodal*, however, they were acquired with different machinery and consequently have different brightness and contrast. For that reason, the input chosen was *multimodal*. This method then outputs a metric and an optimizer. The metric can be one of two types: Mean Square Error (already described) and Mattes MI. This last parameter measures how related two metrics are, so, in this scenario it will describe how related is the fixed and the already transformed moving image. In other words, MI can measure the amount of information that a random variable has about another random variable. MI calculations are done based in entropy between two random variables. Then, considering X and Y as the random variables and H as entropy, the MI between two images is given by equation 3.16.

$$MI(X,Y) = H(X) + H(Y) - H(X,Y) \quad (3.16)$$

In where it comes to the optimizer, there are also two possible outcomes: Regular Step Gradient Descent or One Plus One Evolutionary. These are two commonly used optimizers with the first being a variation of the well-known gradient descent algorithm [87]. The algorithm will adjust the parameters of the image transformation in order for the optimization routine to follow the gradient of the chosen metric, in the direction of an extrema, using a constant step length between evaluations. When there is a change in gradient direction, the optimizer considers that a local extremum was passed and responds by reducing the step-length by a relaxing factor, continuing its search for the best set of parameters until a stopping criterion is reached. The other algorithm is called evolutionary because it starts with a set of parameters – parents - that are mutated (perturbed). If the mutated parameters – child - provide better results, then for the following iteration the child becomes the parent, and the parameters undergo further and more aggressive perturbation. On the other hand, if the parents provide better results, they remain the parents in the next iteration, and the perturbation is less aggressive. This search for the best possible parameters is done until a maximum of the evaluation metric is achieved, or until another stop condition is met.

Finally, in where it comes to transform types, there are four different possibilities that can be considered:

1. *Translation* - Only translational modifications in the *(x,y)* directions can be performed.
2. *Rigid* – Besides translational modifications, this type of transformation also allows rotations.
3. *Similarity* – This transformation is based in translation, rotation, and the possibility of scaling by a factor.
4. *Affine* – Besides all the previously referred transformations is also possible to shear the image.

For each patient, the oldest image was used as reference. All transformation types were used except for translation because it was considered too simple. To evaluate the best transformation for each patient, and consequently the best registration, Pearson's correlation coefficient was computed, through the equation 3.17.

$$p = \sum_{i=1}^{N} \frac{(X-\mu_X)}{\sigma_X} \times \frac{(Y-\mu_Y)}{\sigma_Y} \quad (3.17)$$

With X and Y being the considered images, and μ and σ the mean and standard deviation of the respective images. Then, for each patient, the Pearson correlation for each registration was calculated and the mean Pearson correlation was considered for choosing the best transformation type.

The next chapter is concerned with showing and discussing the results obtained for the different methodologies explained in this chapter.

4. Results and Discussion

Section 4 is concerned with showing the results of the methodologies explained in section 3 and discuss them. Subsection 4.1 is related to pre-processing results, while subsection 4.2 is concerned with feature related topics – extraction and selection. Classification results are shown in section 4.3, and the results obtained with noised and filtered images are presented, and discussed, in section 4.4. Studies concerning area and breast density influence are explored in subsections 4.5 and 4.6, respectively. In subsection 4.7, %PD capability as a predictor is studied. Finally, while subsection 4.8 presents the development of an interactive app for the detection of BC, subsection 4.9 presents and discusses the results obtained for risk analysis.

4.1 – Pre-processing

The pre-processing part of the work consists, as it was mentioned, in label removal, background subtraction and image normalization. The consecutive use of morphological operators proved to be a good approach in the task of label removal, with all the labels being correctly removed and not considered for further analysis. After labels were removed and only breast and background were present, nipple coordinates were found, as the breast boundaries in horizontal direction, and non-essential background was removed. Once every image had its extra background removed, image normalization was performed, to guarantee that the pixel range was the same for every images. Background removal was correctly done across all images and the same can be said about image normalization. It should be noted that when images from the public dataset had some errors in digitization, that might result in the need for a manual boundary definition. An outline of the results obtained across these steps is presented in Figure 4.1**,** with the left image being the original; the next one representing the binary mask that results from the label removal process; followed by the outcome of background subtraction. Image normalization results are not shown once no visual variation is perceived, since pixel relationship within the same image is maintained.

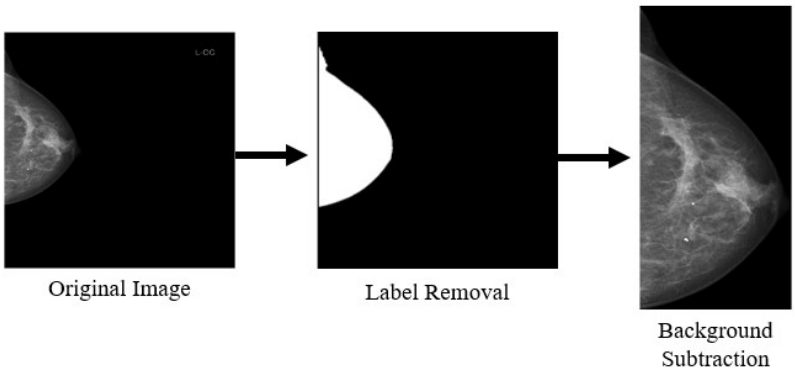

Figure 4.1- Image pre-processing outcome.

For further analysis concerning classifiers performance when dealing with noisy images, gaussian white noise was added to each image of the dataset. The resulting images were then passed through a median filter in order to assess if that method for noise reduction would allow the classifier to present a better discriminative capacity. As described, four different noise levels based on the local variance of the gaussian distribution were considered – 0.01, 0.005, 0.0025, 0.001. The entire

procedure occurred without problems for any image, with the noise distribution seen in Figure 4.2 being as expected.

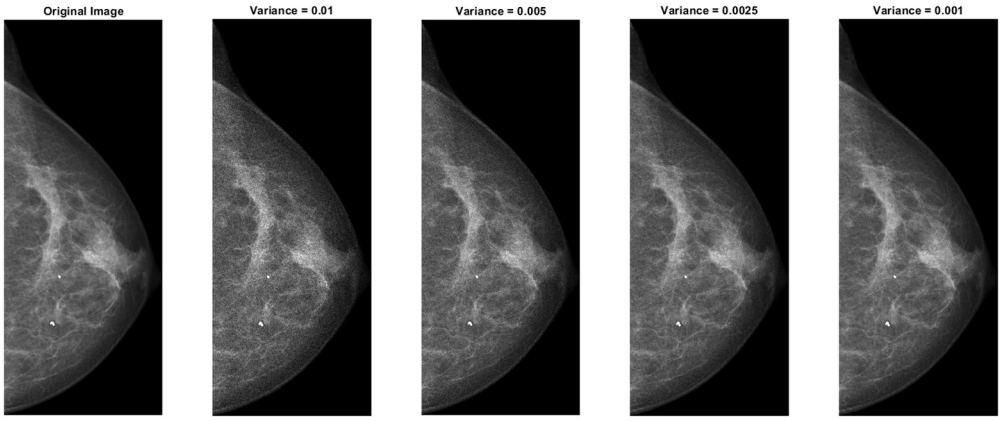

Figure 4.2 - Comparison between original and noisy images, with noise magnitude decreasing from left to right images.

In order to understand median filtering effect, in Figure 4.3 is shown a comparison between an image with the maximum noise level and the resulting filtered image (on the left). The same contrast is made for the image with the smallest noise value (on the right).

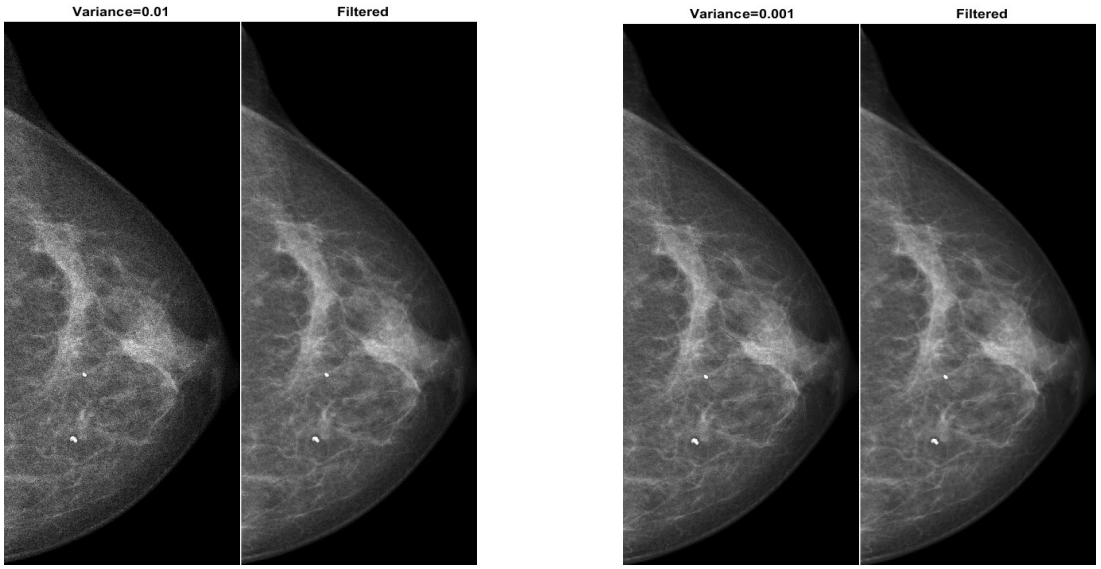

Figure 4.3 - Comparison between noised image and their respective filtered images for the highest (left) and lowest (right) noise value.

As it can be perceived, Median filtering helped to reduce noise that was present in each of the images. Nonetheless, it is possible that a filtering routine, due to its strong nature, when applied to the less noised images, results in loss of images' details. It is important to understand that the study done here takes into account texture patterns, and in the same way that noise can disrupt the interpretation made, the loss of detail in such fine structures might also result in a poor analysis.

To check if the procedures of noise adding and further filtering were done in the correct way, besides considering simple visual cues, the previously presented similarity measures were computed for the different noise values. In Figure 4.4 two plots are shown, each one representing a metric, with

each point being the mean value of that metric across all the images in the dataset, for each noise value.

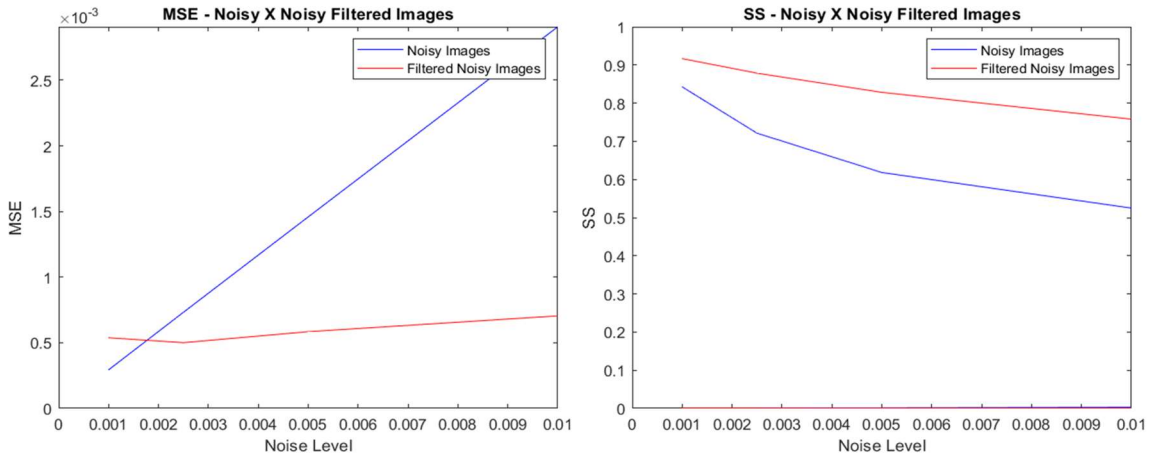

Figure 4.4 - Similarity metrics of the noisy and filtered noisy images, compared with the original images. Mean Squared Error can be seen on the left, while the right plot represents Structural Similarity.

Focusing on the structural similarity measure it is important to remind that it consists of three terms: one related to luminance, other related to contrast and a final one, that it is concerned with the structures in the image. As it can be seen, on the right in Figure 4.4, considering Noisy images, the structural similarity decreases as the level of noise in the image increases. The same trend is observed for the noisy images that were filtered. When comparing the values between these two sets of images, the results are as expected, with the Filtered Images being more structurally similar to the original than the Noised Images.

Nonetheless, it is also important to consider the results related to the mean squared error calculations. As expected, in general, the error for the filtered images is much lower than the one present by the noisy images. However, the results for the lowest noise value seem to contradict the idea that filtered images are *always* more similar to the original image than the noisy images. For a noise value of local variance 0.001, MSE is lower in the noised image than in the filtered image, which could be an indicator that when images with lower noise values are filtered, they might lose some detail, and then become less similar to the original image.

Furthermore, it should also be noted that when comparing the MSE for the noisy image and for the filtered images, the difference between the errors becomes bigger with the increase of noise in the images, probably meaning that images with a great amount of noise benefit more from filtering than images with less noise.

4.2 – Feature Extraction and Feature Selection

Feature Extraction occurred without problems for every set of images and across all features' groups. After the extraction was done through the normalized images, feature selection methodologies were employed, using Orange Software. The results were analyzed in terms of overall and relative AUC values, meaning that not only the absolute AUC value was considered for choosing the best set of features, but also the variation before-after excluding irrelevant features. These AUC values are given by the software using embedded classifiers and do not reflect the results obtained by the later

developed algorithms through MATLAB. K-coefficient variation was also used as a comparative metric to assess model performance between the entire feature set and the relevant features set. The results for the three different criteria in choosing relevant features - Relief-F, Information Gain and Chi-Squared - are presented in the Tables 4.1-4.3. It should be noted that the "entire feature set" is the same for every criterion and, for that reason, the results are the same across the three tables.

Relief-F	Entire Feature Set	Relevant Features	Classifier
AUC Value	0.667	0.718	Decision Tree
	0.828	0.844	SVM
	0.766	0.901	Logistic Regression
K-Coefficient	0.316	0.427	Decision Tree
	0.528	0.517	SVM
	0.38	0.6	Logistic Regression
# Considered Features	88	26	-

Table 4.1 - Feature Selection Results using Relief-F criteria. Comparison of models before and after eliminating redundant features.

Chi-Squared	Entire Feature Set	Relevant Features	Classifier
AUC Value	0.667	0.712	Decision Tree
	0.828	0.808	SVM
	0.766	0.779	Logistic Regression
K-Coefficient	0.316	0.5711	Decision Tree
	0.528	0.497	SVM
	0.38	0.48	Logistic Regression
# Considered Features	88	51	-

Table 4.2 - Feature Selection Results using Chi-Squared criteria. Comparison of models before and after eliminating redundant features.

Information Gain	Entire Feature Set	Relevant Features	Classifier
AUC Value	0.667	0.709	Decision Tree
	0.828	0.829	SVM
	0.766	0.777	Logistic Regression
K-Coefficient	0.316	0.540	Decision Tree
	0.528	0.4545	SVM
	0.38	0.395	Logistic Regression
# Considered Features	88	68	-

Table 4.3 - Feature Selection Results (AUC and K-Coefficient) using Information Gain criteria. Comparison of models before and after eliminating redundant features, for three different classifiers – Decision Tree, SVM and Logistic Regression.

After analyzing each table, it is possible to perceive that after feature exclusion, AUC values increased across all classifiers, with only one exception (SVM Chi-Squared). Further on, an analysis criterion by criterion will be made, both in terms of AUC and K-Coefficient variation.

Considering an AUC evaluation done by [67], the obtained values can be interpreted as seen in Table 4.4. K-coefficient interpretation is present in Table 3.5.

AUC value	Classification Value
0.5 – 0.6	Fail
0.6 – 0.7	Poor
0.7 – 0.8	Fair
0.8 – 0.9	Good
0.9 – 1	Excellent

Table 4.4 - AUC value interpretation, proposed by [67]

For the Relief-F criterion, DT classifier transitioned from a poor performance model to a model with a fair performance. SVM maintained its good performance, despite having a small increase in the absolute AUC value. As for Logistic Regression model, a great variation was observed, with the model changing from a fair to an excellent performance, when eliminating redundant features. Analyzing k-values for the same criterion, Decision Tree passed from a fair to a moderate agreement, SVM classifier maintained its moderate agreement, and Logistic Regression model transitioned from a fair to a near substantial agreement.

Considering the Chi-Squared criterion, the poor performance of the DT algorithm became fair, after feature reduction was applied. SVM model, as it happened for the Relief-F case, kept its good performance, despite a slight decrease in its absolute AUC value. Besides that, Logistic Regression also maintained its fair performance. When assessing k-coefficient values, the results are analogous to what happened with the Relief-F criterion, with exception for the Logistic Regression model. In the Chi-Squared case, the k-coefficient of the Logistic Regression model, that had a fair agreement when considering the entire feature set, showed that there was a moderate agreement between the predicted and the real class labels, with the chosen feature set.

Finally, for the Information Gain criterion, DT algorithm had a poor performance before feature reduction and revealed to have a fair performance after that procedure. As for SVM, the performance remained good after feature selection. Logistic Regression, as it happened for the Chi-Squared criterion, maintained its fair performance. When looking for strength of agreement, DT transitioned from a fair to a moderate agreement, while SVM and Logistic Regression models kept their moderate and fair agreement, respectively.

Information Gain results showed that the variation concerning k-coefficient was approximately non-existing for the Logistic Regression model, with the classifier remaining with a fair strength of agreement, which did not happened for any other classifier, in any criterion. For this reason, and considering its high dimensionality (68 features), Information Gain set was disregarded. Comparing the absolute AUC values of each classifier, for the two remaining criteria, they were similar for Decision Trees, while for SVM the results obtained by Relief-F were relatively higher. For the Logistic Regression model, the value was substantially different between criteria, with the result from Relief-F being higher than the one obtained with the Chi-Squared criterion. The variations of the k-coefficient were similar for both feature sets, despite the results obtained for Logistic Regression, where the Relief-F set achieves a higher strength of agreement.

In spite of the Logistic Regression results, the two sets appear to have similar potential for further algorithm development. However, considering the Logistic Regression results, and since the

Relief-F set is considerably simpler than the Chi-Squared set (26 vs. 51 features), the set obtained through this criterion was the chosen one.

Table 4.5 shows, by number, the features chosen by this criterion – for feature description see Tables 3.1, 3.2, and 3.3 – as their division across the previously introduced feature groups.

Relief-F	Intensity-Based	Run-Length	Co-Occurrence
Features Chosen	2	23, 26	6, 7, 14, 18, 21 - (0º, 45º, 90º, 135º); 15 (0º, 90º); 16 (0º)
# Features	1	2	23

Table 4.5 - Features Selected through the Relief-F criterion

4.3 – Algorithm Development and Classification

The entire dataset was reduced, with data concerning to features that were not selected being eliminated. With data division into training and testing done, model construction could start. Algorithm optimization, as explained in section 3, occurred during the training part of the work, with parameters being tuned recurring to cross-validation (5 or 10-fold) methodologies. Ten different variations of optimization options were considered, which resulted in ten different trained models for each classifier – except for Logistic Regression, once again, due to MATLAB constraints. For each trained and optimized classifier, its performance was assessed using four sets of images that the algorithm had never seen before, and that included both cancer and normal (or benign) cases. The different variations in terms of options chosen for: Optimizer, Acquisition Function, Maximum allowed iterations, and Validation technique for hyperparameter tuning, were presented in Table 3.6.

The choosing of the optimization options to be consider was done during a test round, where between all the examined Acquisition functions, the two presented in Table 3.6 appeared to be the most promising. In terms of Maximum Allowed iterations, both Bayesian and Random-Search achieved satisfactory results with the default value (30) and no significant increase was observed when this value was raised. On the other hand, Grid-Search performed poorly with 30 iterations, which explains the choosing of the values 100 and 500 for this parameter, that seemed the most encouraging during the test round. For Validation, the option chosen was cross-validation with both 10-fold and 5-fold, used for hyperparameter tuning. Once again, each classifier – SVM, DT, DA, LR - was tested with four different test sets and, for this reason, each classifier variation presents four different AUC value, one for each test set.

Tables 4.6, 4.7 ,and 4.8 present the results obtained for SVM, DT and DA, respectively. The Logistic Regression model used was embedded in MATLAB software and did not allow any optimization options. For that reason, results relate to only one variation – Table 4.9.

SVM Variation	AUC Test set 1	AUC Test set 2	AUC Test set 3	AUC Test set 4
1	0.836	0.823	0.816	0.762
2	0.865	0.829	0.868	0.845
3	0.875	0.865	0.860	0.869

4	0.819	0.872	0.838	0.827
5	0.848	0.855	0.860	0.857
6	0.848	0.855	0.860	0.857
7	0.855	0.829	0.868	0.845
8	0.819	0.872	0.838	0.827
9	0.848	0.848	0.860	0.863
10	0.848	0.812	0.848	0.857

Table 4.6 - SVM variations, in terms of AUC, for different test sets (1-4), across the ten classifier variations

.

Decision Tree Variation	AUC Test set 1	AUC Test set 2	AUC Test set 3	AUC Test Set 4
1	0.711	0.711	0.789	0.643
2	0.720	0.720	0.736	0.601
3	0.737	0.737	0.789	0.654
4	0.711	0.711	0.789	0.642
5	0.687	0.687	0.648	0.565
6	0.711	0.711	0.789	0.642
7	0.687	0.687	0.648	0.565
8	0.730	0.730	0.753	0.625
9	0.737	0.737	0.789	0.655
10	0.737	0.737	0.789	0.655

Table 4.7 - Decision Tree variations, in terms of AUC, for different test sets (1-4), across the ten classifier variations.

Discriminant Analysis Variation	AUC Test set 1	AUC Test set 2	AUC Test set 3	AUC Test set 4
1	0.848	0.848	0.855	0.813
2	0.848	0.848	0.855	0.813
3	0.848	0.848	0.855	0.813
4	0.822	0.822	0.860	0.822
5	0.848	0.848	0.819	0.780
6	0.839	0.839	0.843	0.846
7	0.848	0.848	0.819	0.780
8	0.839	0.839	0.843	0.846
9	0.796	0.796	0.738	0.722
10	0.848	0.848	0.819	0.780

Table 4.8 - Discriminant Analysis variations, in terms of AUC, for different test sets (1-4), across the ten classifier variations.

LR Variation	AUC Test set 1	AUC Test set 2	AUC Test set 3	AUC Test set 4
1	0.8194	0.8080	0.8194	0.8130

Table 4.9 - AUC value, calculated on the test sets for Logistic Regression, which due to MATLAB constraints only has one variation.

When comparing the overall AUC of SVM classifiers against the results obtained for Decision Tree, it can be seen that they are substantially higher in the SVM. Actually, there is only one variation – classifier 1 in test set 4 - with an AUC below 0.8, across the different test sets. On the other hand, Decision Tree trained classifiers do not present any value above 0.8, in any of the four test set analyzed. These considerations allowed to disregard Decision Tree classifiers for the decision of which classifier to choose.

Considering now the SVM and the DA results, the AUC values appear to be more equal between classifiers. When analyzing the performance of the ten variations for each classifier, across test sets, it is possible to see that the mean AUC of the SVM classifiers is higher than the one obtained for DA for every test set: 0.8461 vs. 0.8384, for test set 1; 0.8463 vs. 0.8390, for test set 2; 0.8516 vs. 0.8303, for test set 3; and 0.8409 vs. 0.8015 for test set 4. Besides that, for each test set, the highest AUC across variations is always found in the SVM results. Finally, the result obtained for the Logistic Regression Classifier is lower than the mean AUC for the SVM classifiers across different test sets.

For the presented reasons it becomes clear that the classifier to be considered should be an SVM classifier. Looking only for the results obtained with these algorithms, the higher AUC value was obtained in variation 3, with test set 1. However, the results obtained for variations 4 and 8 with test set 2, are not far from that value. To validate that the classifier with the highest AUC was better, k-coefficient was calculated for the two classifiers. Evaluating the results in terms of correct and incorrect classifications made in test set 1, it was found that the higher AUC classifier presents a k-coefficient of 0.795, achieving a strength of agreement in the borderline between substantial and almost perfect. On the other hand, the classifier with an AUC of 0.872 had k-value of 0.770.

Different optimization variations (1 to 10) resulted in different hyperparameters selected – in the case of SVM, different Kernel Functions or Box Constraint values. The chosen variation (variation 3) which used a Bayesian Optimizer, Expected Improvement acquisition function, and 10-fold cross-validation, resulted in a Box Constraint = 3.6997; a Kernel Scale = 1; and a Kernel Function = linear. This was the SVM algorithm was the classifier used for further studies, and the test set that produced this AUC result (test set 1) was the one explored, for example, for studies concerning breast area and image noise bias in classification.

4.4 – Noise Results

The chosen SVM classifier was tested using the noised and filtered variations of the images included in test set 1. For that reason, eight different AUC results were obtained. Four of them represent the results of the classifier tested with noised images, and the remaining four results correspond to the testing of the classifier using the same noised images after passing them through a median filter. Table 4.10 shows the absolute AUC values obtained for each of the described situations, while the plots present in Figure 4.5 represent the same results but as a visual variation across different noise values for noisy images and filtered imaged.

	Standard Deviation of Gaussian White Noise			
	0.001	0.0025	0.005	0.01
Noised Images	0.765	0.637	0.469	0.475
Noised Images after Filtering	0.500	0.526	0.554	0.754

Table 4.10 - AUC variation across different Noise Levels, for noised images before and after filtering, from the lowest (left) to the highest (right) noise level.

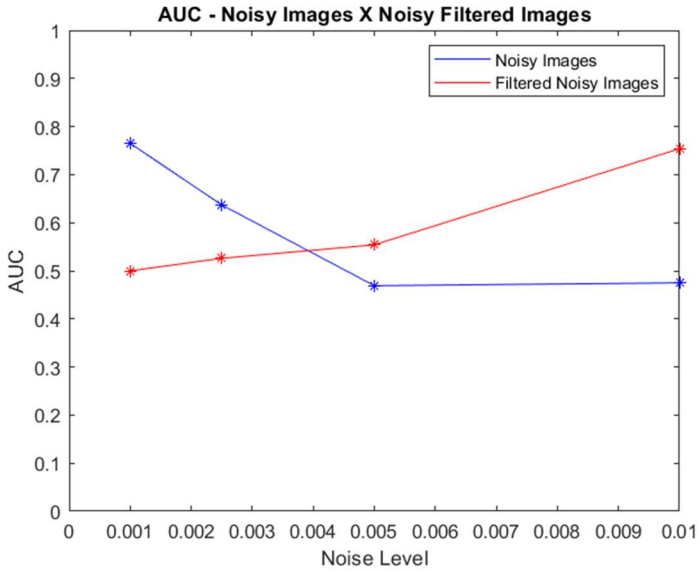

Figure 4.5 - AUC variation for different noise values, for noised images before – red - and after – red – filtering.

The results from the classifier testing using the noised images showed what was expected at first: the performance decreases with an increase in noise level. Nonetheless, even when noise is present in the images, although in small amounts, the classifier preserves a fair, nearly good, performance, with an AUC value of 0.765. With a noise value of 0.0025 the classification drops to a poor performance, and from that level up, the classifier fails to make valuable predictions.

When the images are filtered, noise levels 0.001 to 0.005 produce no significant results in terms of classification. For the lowest noise magnitude, the filtering routine makes the classifier to have a poorer performance than when tested with the noised images. This result may be related to what was observed in Figure 4.4, where for the smaller noise level, the filtered images had a MSE superior to the noised images.

Nonetheless, when looking for the higher noise value, the noisy images presented an AUC below 0.5, indicating no discriminative capacity at all. However, when these images were filtered, the

AUC achieved a value of 0.754. This result might indicate that a filtering routine, with a median filter, might improve image quality without loss of detail, for images with high noise values.

4.5 – Breast Area Influence in Classification

The work done in this thesis was different from what was presented in the literature review section, once the entire breast is used for image analysis, instead of using a pre-defined ROI, with fixed dimensions. Here, not only image size was not fixed, as the same can be said about breast dimensions. The goal of this part of the study was to verify if overall Breast Sizes, or the percent area of the image occupied by the breast, biased the algorithm in terms of classification made. The results (concerning Test set 1) for the overall Breast area, and percent area of the image occupied by breast tissue for: healthy images from Luz dataset; benign images from the DDSM dataset; and cancer images from the DDSM dataset are shown in Table 4.11. Figure 4.6, serving as a visual aid, depicts the variation of Breast Area and % area occupied by breast tissue for 15 randomly selected cases of each of the three previously defined sets of images.

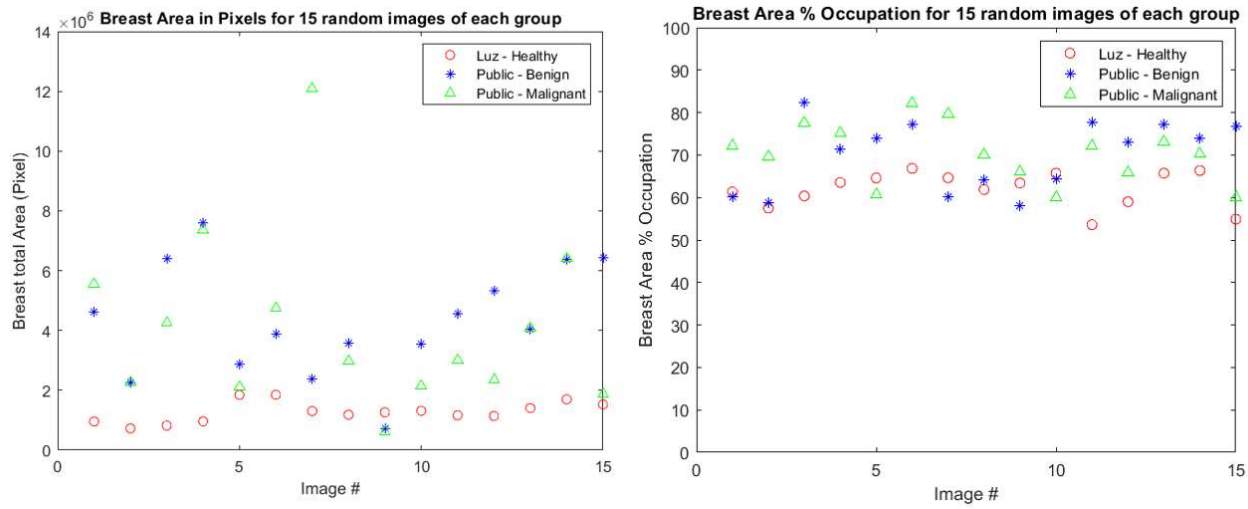

Figure 4.6 – Total Breast Area, on the left, across the 15 randomly selected images, for each group. On the right it is represented the area occupied by the breast, in percentage, the 15 randomly selected images, for each group.

	Mean Breast Area (pixels)	Mean Area Occupied by the Breast (%)
Luz data	1 318 133.365	63.67
Healthy data – Public dataset	4 459 686.328	69.73
Cancer data – Public dataset	5 475 580.55	69.75

Table 4.11- Mean Total Area of the image and Mean Percent Area occupied by the breast, across test set 1.

Looking for Figure 4.6, it can be perceived that both in terms of absolute breast area, and in terms of area occupied by the breast, the images that come from the Luz dataset are substantially smaller than the ones that were retrieved from the DDSM dataset. This observation is corroborated by the results presented in Table 4.11, where Luz data has the smallest results both in terms of Mean Breast Area and Mean Percent Area Occupied by the breast.

The distribution of the public dataset cases seems random, despite concerning to cancer (malignant) or healthy (benign) cases, with both the highest and lowest cases in terms of area belonging to the cancer class. These observations indicate that there is no clear pattern, in terms of breast area, that distinguish the healthy and the cancer cases retrieved from the public dataset. When searching for the misclassified point in the test set 1, it was possible to see that these misclassifications occurred for both cancer and non-cancer cases across different values of breast area. There were cancer cases with breast area equal or even lower than the Luz being correctly classified, and the same can be said about cancer cases with breast area values higher/equal than the area presented by healthy cases of the public dataset. The fact that there were correctly classified cases, healthy and cancer, across the entire range of area values, and that the same occurred for misclassified points, gives confidence that breast area is not influencing classification.

To give more confidence to this idea, a correlation coefficient between Breast Area and output class was calculated. The same coefficient was computed between the % area occupied by breast tissue and the output class. The obtained values are presented in Table 4.12.

.

	Output class
Breast Area	0.223
% Breast Area Occupation	0.123

Table 4.12 - Correlation Coefficient between a) Image Area and Output Class, and b) % Area Occupied by Breast Tissue and Output class

The results obtained provide a clue that there is not a clear trend between the classification being made and the area of the images, or the breast size. The fact that the correlation achieves results so low, associated with the previous analysis made to the test set gives confidence when drawing these conclusions.

4.6 – Breast Density and Classification

Increased Breast Density, as explored in the first sections of this work, is not only a risk factor of BC as may also be an obstacle in finding lesions through mammography. BC detection, done by an expert professional, becomes more difficult as breast density increases, with the accuracy rate of cancer detection decreasing. For that reason, it is important to very if the develop classifier accompanies this trend, or if the performance of the algorithm is independent of breast density.

The classifier was tested with five different test sets for each density class, in order to avoid bias concerning the images chosen. The AUC results for density classes 1 to 3, are present in Table 4.13, for each density test set.

	BI-RADS 1	BI-RADS 2	BI-RADS 3
Density set 1	0.833	0.9643	0.8929
Density set 2	0.875	0.9333	0.9643
Density set 3	0.850	0.8571	0.8571
Density set 4	0.863	0.9583	0.8846
Density set 5	0.900	0.9231	0.9286
Mean AUC	0.8642	0.9272	0.9055

Table 4.13 -AUCs, for different breast density classes (BI-RADS 1-3), across five different test sets (Density set 1-5).

The results for BI-RADS 1, the lowest density class, seem positive, with AUCs near the value obtained for the classifier when tested with test set 1. When looking for the BI-RADS 2 class, the results seem even more positive to what happened for the first class of breast density. The results for BI-RADS 3 revealed to be extremely positive, with some of the test sets resulting in AUCs over 0.9. However, in mean terms, there was a small decrease when comparing these AUC values with ones obtained for BI-RADS 2.

The tendency observed in mean AUC, in Table 4.13 is not as linear as it was expected at first. There is an increase in classifier's performance when breast density shifts from 1 to 2. Then, as expected, there is a decrease in the performance when density increases again. This occurs, probably, to the low dimensionality of the test set used to assess the classifiers performance with the BI-RADS 1 test set. Having a small dimension, a misclassified case has a greater impact than a misclassified case in the BI-RADS 2 and BI-RADS 3 test set.

As it can be perceived, the results concerning BI-RADS 2 images proved to be high across different test sets, with a higher performance than the one obtained for BI-RADS 1 but lower than BI-RADS 3. Even though a decrease is observed when breast density increases for class 3, it should be noted that the general classifier performance is not deeply affected. This fact can be corroborated by the mean AUC of the classifier when being tested with a test set composed of BI-RADS 3 images (AUC=0.9055). Finally, for BI-RADS 4, the public dataset provided only 1 case of a healthy subject and for that reason, a correct analysis with AUC calculation could not be conducted. Nonetheless, the results obtained for the three different density classes are promising and appear to indicate that the classifier is robust to breast density, in the task of differentiating cancer from non-cancer patients.

4.7 – Percent Mammographic Density as a Cancer Predictor

The effect of %PD in the discrimination of cancer and healthy patients has been explored in the state-of-the-art section. The study concerning the effect of %PD done here, consisted of training new classifiers, with the same images previously used in the training set, and tested with the images of test set 1. The results of the novel trained classifiers are present in Table 4.14, 4.15 and 4.16. For comparison purposes, in the same tables, the results obtained with test set 1, in the original classifiers, are repeated.

SVM Variation	AUC – Original Classifier	AUC – Classifier with %PD
1	0.836	0.849
2	0.865	0.849
3	0.875	0.829
4	0.819	0.865
5	0.848	0.829
6	0.848	0.829
7	0.855	0.829
8	0.819	0.829
9	0.848	0.829
10	0.848	0.848
Mean	0.8461	0.8386

Table 4.14 - SVM: AUC comparison between original classifier and one that incorporates %PD, for the ten classifier variations.

Decision Tree Variation	AUC – Original Classifier	AUC – Classifier with %PD
1	0.711	0.764
2	0.720	0.747
3	0.737	0.704
4	0.711	0.711
5	0.688	0.747
6	0.711	0.688
7	0.688	0.701
8	0.730	0.704
9	0.737	0.688
10	0.737	0.763
Mean	0.717	0.722

Table 4.15 – Decision Tree: AUC comparison between original classifier and one that incorporates %PD, for the ten classifier variations.

Discriminant Analysis Variation	AUC – Original Classifier	AUC – Classifier with %PD
1	0.848	0.839
2	0.848	0.875
3	0.848	0.865
4	0.822	0.865
5	0.848	0.796
6	0.839	0.829
7	0.848	0.796
8	0.839	0.829
9	0.796	0.839
10	0.848	0.796
Mean	0.838	0.833

Table 4.16 – Discriminant Analysis: AUC comparison between original classifier and one that incorporates %PD, for the ten classifier variations.

Making a comparison between the before-and-after %PD adding to the set of features, in both SVM and DA the mean AUC suffers a slight drop in its value, while with the Decision Tree Classifiers, a slight increase can be observed. These results might indicate that the information given by %PD does not contribute widely for cancer detection. Having a deeper look at each classifier, for SVM the lowest AUC for the original test set was 0.819 , while for the set that contained %PD was of 0.829. This result could, in theory, indicate that %PD contributes with important information for the discrimination between healthy and cancer cases; however, the result is counterbalanced by the higher AUC value obtained – 0.875 for the original test set and 0.865 for the test set with %PD.

For Decision Tree trained classifiers, there is an increase in the mean AUC value obtained across variations; and, although the minimum AUC value is the same for both classifier groups, the maximum value of the set that considers %PD as a descriptor is of 0.764, while for the original test set is of 0.737. Finally, for DA, as it was said, the mean AUC value obtained decreases when %PD is added to the features set. The lowest value obtained is the same for both the original classifier and the classifier that considers mammographic density (0.796), However, the highest value is obtained for

the classifier that has %PD as a feature. This highest value is of 0.875, which is the same value of AUC as the one obtained by the chosen classifier (SVM – variation 3). In this scenario, the question of which classifier – the DA with %PD or the original – was the best could arise, however, since the simplest model is usually the one that best describes the phenomenon in study, the third variation of the original SVM remains as the chosen one.

Considering the results described here, one might understand that for classifiers that already have a good-to-excellent performance, in general, %PD does not add significant information, as it can be seen by the mean AUC values of SVM and DA. On the other hand, for fair-to-good classifiers, as it is the case of Decision Trees, in the scope of this work, %PD might be a feature that contributes to an increase in the discriminative performance of the classifiers.

Given that, and considering the results concerning mean , and minimum and maximum AUCs, %PD was not considered a relevant feature. For that reason, the variation 3 of the firstly trained SVM classifier was not disregarded.

4.8 – Development of an Application for Automatic Cancer Detection.

The development of the interactive application was also done in MATLAB and aimed to be as simple as possible, so that it could be used easily. The goal of the application is to classify images as cancer or healthy images, and it can be used without needing a connection to the Internet. A brief explanation of the interface is given below, with the visual aid of Figures 4.7 to 4.12.

When the user clicks on the application icon, the menu present in Figure 4.7 pops up.

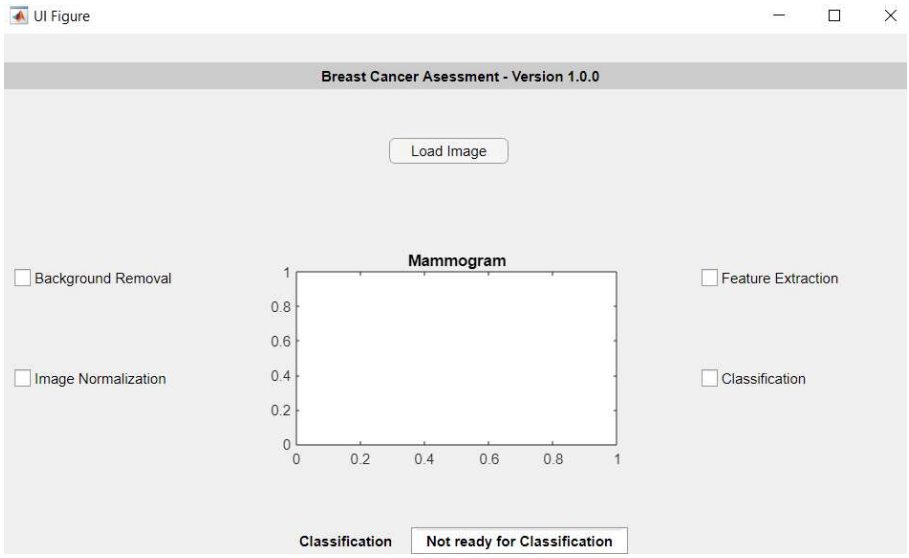

Figure 4.7 - Interactive Application outline.

The top button allows the user to navigate to different locations and folder in order for them to choose the image that they want to analyze. On the left-hand side of the application, there are two checkboxes related to image pre-processing. The top one is related to background removal. After loading the image and checking the "Background Removal" box, a display of the cropped image is shown in the center panel. To grant that the background elimination process occurs as correct as possible, a dialog box appears, where the user has the possibility to state if they are satisfied or not with background removal. If they are not satisfied, the image will pop-up in another window and the user has the opportunity to, interactively, define a rectangular region to crop the image. This process is exemplified in Figures 4.8, 4.9 and 4.10.

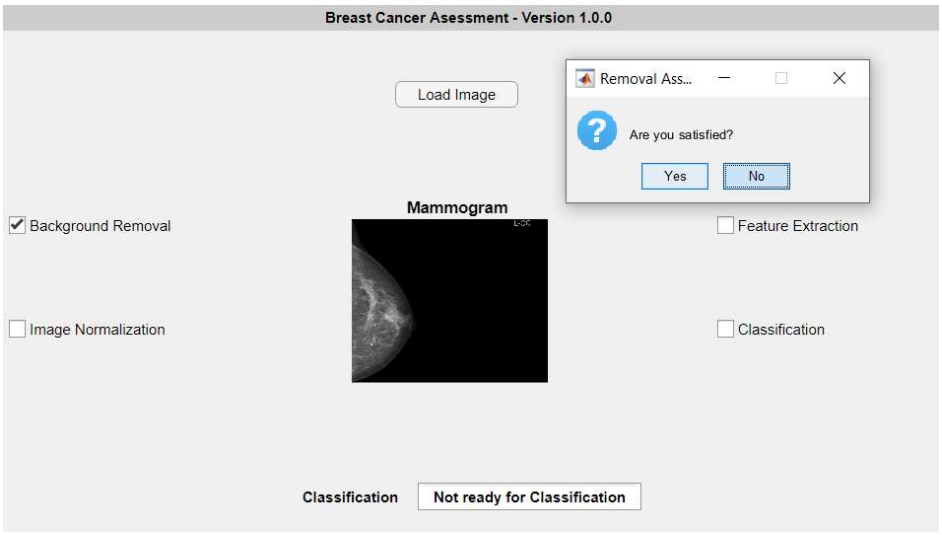

Figure 4.8 - Bad automatic Background Removal and Dialog Box to check for the quality of the process.

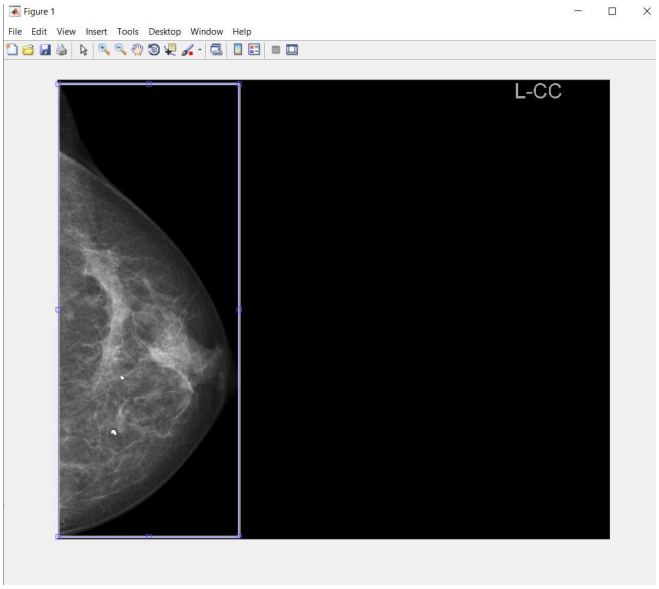

Figure 4.9 - Image that pops-up and interactive background removal.

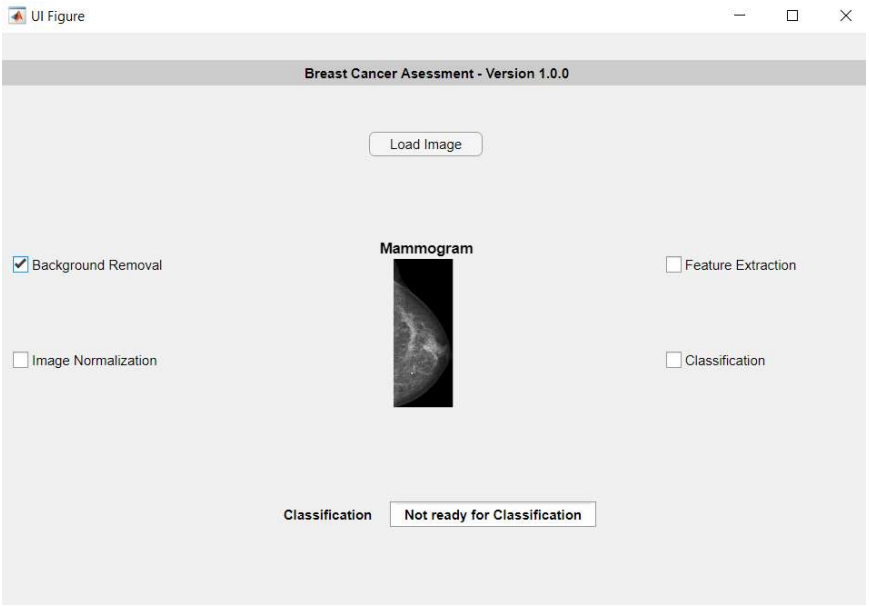

Figure 4.10 - Application menu after interactive background removal.

After this is done, Image Normalization checkbox can be clicked, and the pre-processing steps are completed.

Looking for the right side of the interface, one must first look to the feature extraction checkbox. By clicking on it, features will be extracted from the pre-processed image, and since it is a slower procedure, a process bar is shown in the menu, as depicted in Figure 4.11.

After feature extraction procedure is complete, the textbox in the bottom of the menu becomes "Ready for Classification".

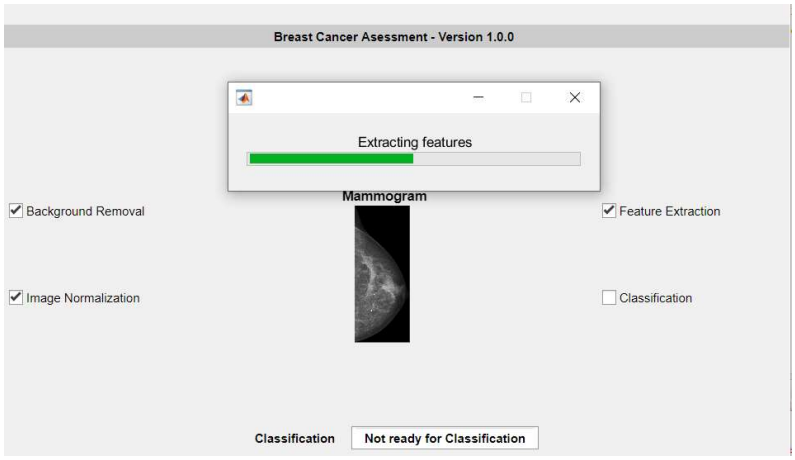

Figure 4.11 - Feature Extraction Procedure accompanied by the process bar.

Finally, the image is ready for classification, what can be done by clicking on the classification checkbox. If the image belongs to a healthy woman, the classification appears in green, if not, the textbox becomes red. Figure 4.12 depicts this procedure for a woman that does not have the disease.

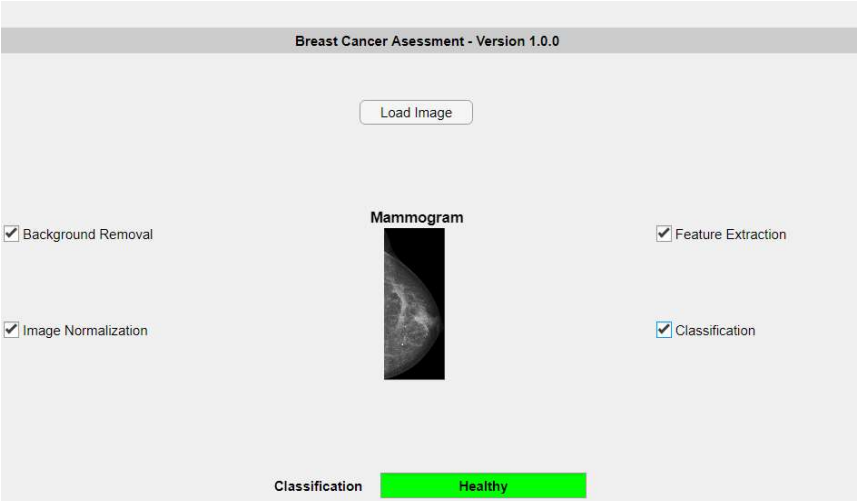

Figure 4.12 - Classification Results, with the green color appearing since it is a healthy case.

4.9 – Risk Assessment

For risk assessment, first, it was important to register all the images from the same patient. The registration accuracy was measured with the Pearson Correlation coefficient, which also allowed to choose the image transformation that was best for each patient. Figures 4.13, 4.14 and 4.15 represent the different transformations – Rigid, Affine and Similarity, respectively - considered for a specific patient across the years. For each image, the correlation coefficient is displayed. The metric and optimizer chosen were MI and One Plus One Evolutionary, respectively

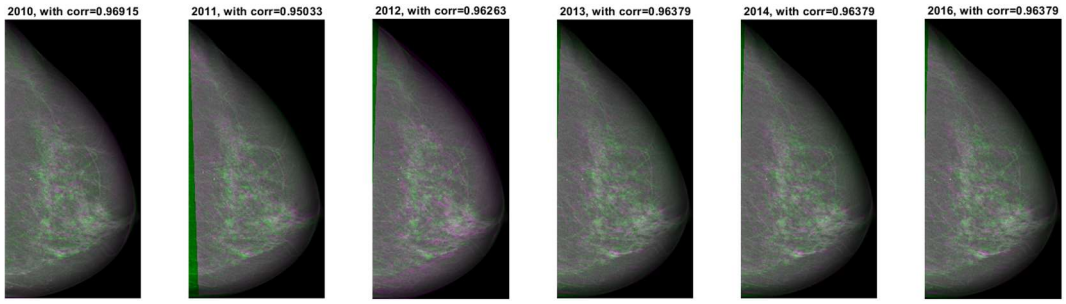

*Figure 4.13 - Image Registration results, using 2009 as fixed image, with a **Rigid** Transformation. Correlation Coefficients presented above the registered images.*

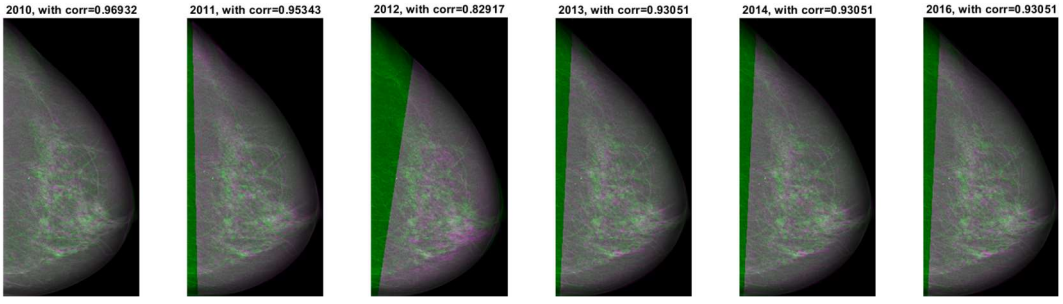

*Figure 4.14 - Image Registration results, using 2009 as fixed image, with a **Affine** Transformation. Correlation Coefficients presented above the registered images.*

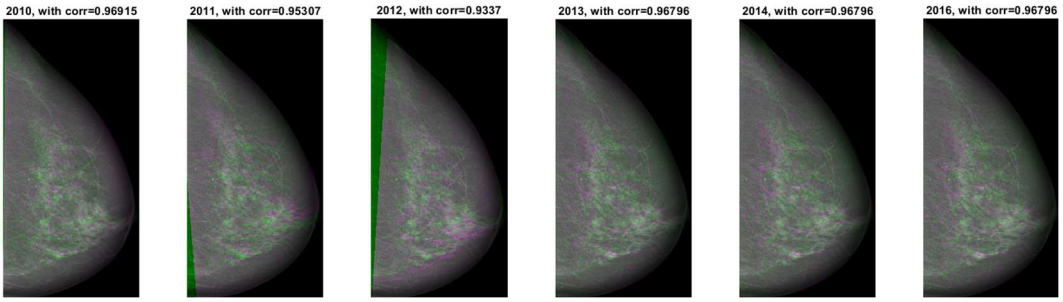

*Figure 4.15 - Image Registration results, using 2009 as fixed image, with a **Similarity** Transformation. Correlation Coefficients presented above the registered images.*

Visually, it can be perceived that for the case of this patient, the affine transformation is the one that produces poorer results, which is corroborated by the correlation values presented – the smallest value (0.82917) is present in a registration done with an Affine transformation. To decide what was the best transformation, a mean correlation value was computed. Rigid transformation produced a mean correlation coefficient of approximately 0.962, while the Similarity transformation had a mean value of nearly 0,960. For that reason, for this patient, Rigid transformation was chosen.

This procedure was done for every patient, and throughout different patients, different transformations were chosen as the best, meaning that there was not a clearly superior transformation across patients.

Once registration was correctly performed for every patient, feature extraction explained in section 3 occurred without any problems. The following part of the work was related to the analysis of feature variation across the years for each patient.

Feature visual analysis was done by looking for plots like the ones present in Figures 4.16 and 4.17 (Feature Sum Variance). The analysis of this of these plots will be done group-by-group, and, inside each group, in an image-by-image basis. As it was described in the beginning of section 3, images from 2007 and 2008 were acquired with different machinery. For that reason, they were not considered for previous studies, however, here, they were considered. It can be noted, looking for any plot that contains images from these years, that there is always a big discrepancy between feature values in these and the subsequent years. For this reason, for plots where there are feature values after 2008, the analysis performed does not consider feature values for images before 2009.

Once again, the goal here was to assess feature fluctuations more than to verify what were in fact the absolute values that features assumed in sequential years. Besides “Sum Variance”, other features had the same behavior, both in Co-occurrence (Contrast) and Run-Length (Gray-Level Nonuniformity, Run Percentage, etc.).

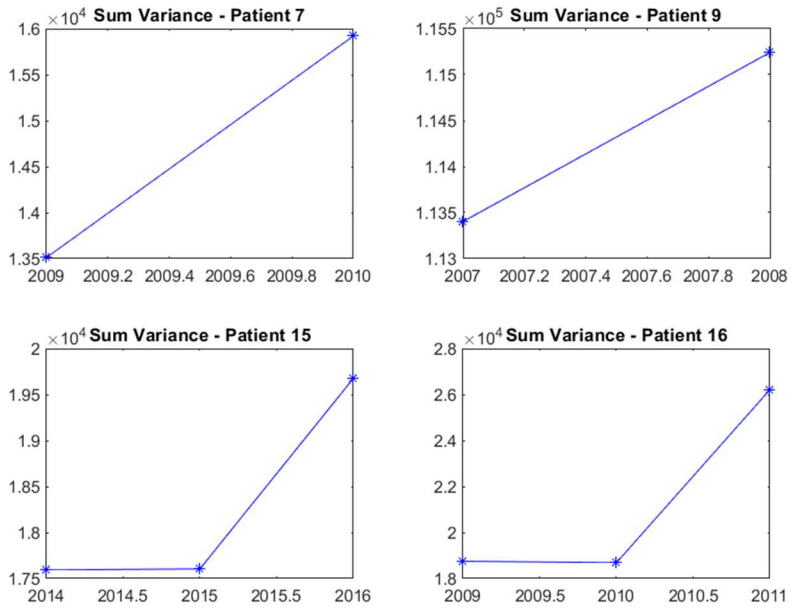

Figure 4.16 - Sum Variance feature value variation across years, for different cancer cases (Patients 7,9,15,16), until the year immediately before cancer diagnosis

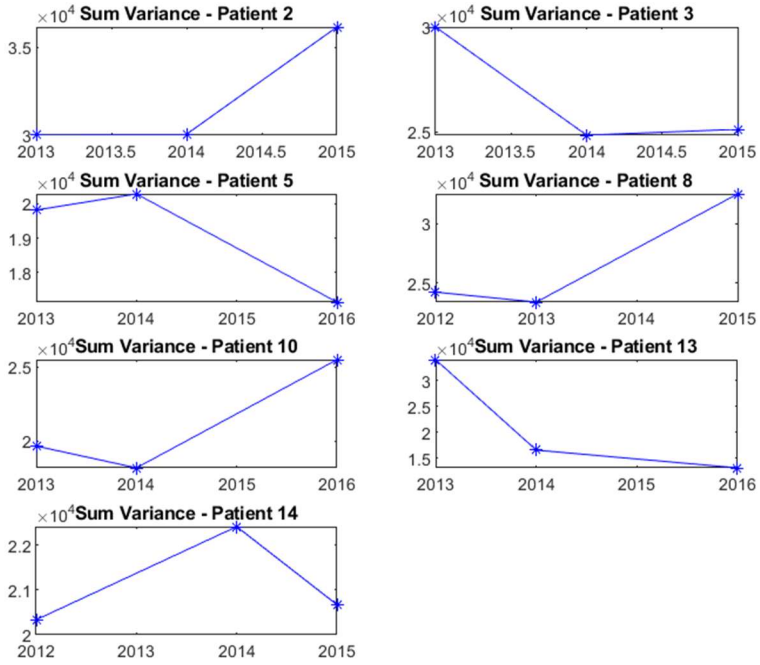

Figure 4.17 - Sum Variance feature value variation across years, for different healthy cases (Patients 2,3,5,8,10,13,14), until the year immediately before cancer diagnosis.

Starting in the cancer group – seen in Figure 4.16 – and beginning by patient 7, it can be seen that between that there was an increase in feature value from 2009 to 2010. This patient would have cancer being diagnosed in 2011, which means that in the year immediately before cancer was diagnosed, this feature went up. When analyzing patient 9, this tendency is maintained. For patients 15 and 16, it is possible to see that between sequential years: 2014/2015 in patient 15, and 2009/2010 for patient 16, this feature maintained its value nearly equal. However, the value went up in the year

immediately before cancer diagnosis as it happened to the other two patients. The analysis conducted in the cancer group indicated that this feature could have a positive impact when assessing breast risk. However, the healthy group should be addressed first. It should be noted that for the healthy group, there were more than three images, but only three results are shown to allow a fair comparison with the cancer group. Moreover, the feature value presented in the first year for each case of the healthy group does not reflect the absolute feature value of that year, but rather a mean of the values that the feature assumed in previous years. The use of this approach allows not only visual comparison to be more direct, having approximately the same number of values in both cancer and healthy cases, as also the contributions from all years are being considered.

For patient 2, from the first year to the second the feature value appears to remain constant, with little to no variation. However, from 2014 to 2015, there is an abrupt increase in feature value. The behavior present by patient 2 is very similar to what was observed for cancer patients (15 and 16). When looking for patient 3, it is possible to verify that from the first mammogram to the following, there was a diminishing in the feature value, that remained constant in the last year. This behavior is nothing like the ones seen for cancer patients, once not only does the feature not increase (significantly), as it decreases. Patient 5 started off with what seems a negligibly increase. This increase was rapidly counterbalanced by a great decrease in feature value from 2014 to 2016. Although different from what was seen for patient 3, the feature value remains approximately constant and then decreases, no notable increase was seen for this healthy patient. On the other hand, patient 8 had an analogous variation to what was seen for patient 2, having a controlled value between the first and second year, and then seeing an abrupt increase in feature value between the last two mammograms. Patient 10, although with a more noticeable decrease from the first to the second year, also had an abrupt increase between the last two years. The analysis that was done for patient 3 can be repeated for patient 13, although a slight decrease can be note between the last two years. Finally, patient 14 has an abrupt increase from the first to the second-year mammogram, being then counterbalanced, with a decrease from the second to the last year.

Although a clear tendency was seen in the cancer group, with an increase in feature value in the mammogram immediately before cancer was detected, the same cannot be said about the healthy group. Patients 3, 5 and 13 have a behavior clearly different from what happens to cancer cases, with the feature values being constant and/or decreasing. On the other hand, patients 2,8, and 10 seem to assume a tendency similar to the one presented by cancer cases. Finally, patient 14 has a behavior different from the remaining patient set, having an abrupt increase and decrease.

The cases of patients 2,8, and 10 were investigated by a responsible of Hospital da Luz and it was found that patients 2 and 8 had, in fact, later developed breast lesions. On the other hand, patient 10 had underwent breast surgery during the considered time frame, which could indicate that feature variation (the initial decrease and the abrupt increase) might be referent to breast modifications related to breast surgery. Patient 14, as it happened for patients 3,5, and 13, remained healthy.

In summary, the clear pattern that was verified firstly for cancer cases was not verified for all healthy cases, which at first could be a good indicator, once there was a clear difference between groups. However, having a deep look, three of the seven healthy cases present a behavior very similar to the one observed for the “high-risk” group. This fact led to the exploration of these three cases, which allowed to verify that there were in fact significant breast modifications. The process by which this occurred gives confidence that feature extraction and evaluation across years might be a good method to be applied when assessing BC risk. However, these results should be interpreted with caution: as it could be seen, the feature fluctuations did not only increased when cancer was present, but also when parenchyma modifications occurred. So, this type of algorithms should be used to aid the diagnostic (or the risk assessment) and not to substitute the established methods.

In what concerns to the computed metrics to accompany the visual feature fluctuation,

absolute difference and slope calculations concerning features present a confirmation of what it is seen in the plots. The rationale behind developing structural similarity metrics was related to the fact that as the images progress to a cancerous stage, they begin to be more different between each other. However, this measure, besides structure, also takes into consideration luminance and contrast measure, parenchymal changes is not the only factor being considered. Besides that, factors other than cancer could be contributing for the discrepancies in structural similarity. Given that, this measure did not provide any relevant results.

The idea of computing these metrics was creating new features that described what could be seen in terms of feature fluctuations and later fed them to ML algorithms. Once again, given the dataset small dimensionality, this was not possible. Still referring to the options taken in subsection 3.6, scenario 2 was chosen over scenario 1 in order to unveil patterns in feature fluctuations instead of just considering 1-2 years before cancer diagnosis. However, as it can be seen in the “high-risk” group, significant feature increase occurs only in the mammogram immediately before cancer identification. This fact gives strength to future applications in BC risk prediction using not only feature fluctuation as also the approach referred in scenario 1, subsection 3.6.

5. Conclusions and Future Work

Considering the goals presented in subsection 1.9, and pondering the results obtained in section 4, in general, this work served its purposes. First of all, in terms of image pre-processing, background removal was correctly achieved, allowing an analysis that did not contained irrelevant pixels – background and labels. As it can be seen in the figures and tables of section 4.1, noise adding and consequently noise filtering occurred without a problem, and the resulting images are as expected – the higher the noise, the further away they are from the original image.

The features explained in section 1.7 were correctly extracted from every image – original, noised, and filtered. It is important to note the innovative methodology used: instead of extracting features from previously defined suspicious regions, the features were extracted from the entire breast. This methodology increases the difficulty in classification because there is more tissue being characterized. In some cases, there is even healthy tissue being considered in the cancer cases, which could led the classifier into an incapability of differentiating cancer and healthy cases. However, on the other hand, the procedure used allows for a more automatic classification, without the need of the time-consuming laborious work of looking for suspicious regions. Given that, the procedure becomes more automatic, faster, and less prone to human error.

Feature extraction procedure led to feature selection, done in the original images. The criterion used for this feature selection methodology, Relief-F, gives a score of how well a feature can discriminate the two classes, and it was this score that was used to rank the features, and consequently remove irrelevant ones. The comparison of AUC and k-coefficient between the entire feature set and the relevant feature set values proved what was expected, i.e., some of the extracted features were not relevant for the considered task and, for that reason, were eliminated from the feature set – 88 extracted features vs. 26 relevant features. For that reason, as it happened to the pre-processing stage, feature extraction and selection was successful.

Concerning the classification phase, ten different variations of each classifier algorithm were trained, each with different optimization options (Optimizer, Acquisition Function, etc.). These optimized classifiers were then tested with four different test sets, to avoid bias in the test set chosen. After that, mean AUC values obtained were used to disregard non-relevant relevant classifiers. It became clear through this analysis that Support Vector Machine was the superior algorithm, and in order to choose the best SVM algorithm, k-coefficient measure was calculated for the ones that achieved the highest AUC value. This process resulted in the choosing of a classifier with a AUC value of 0.875.

The chosen classifier was tested with the noised test set and the filtered test set. When looking to the noised image results, it was possible to see that noise presence, even in low levels, results in a lower AUC when compared to the original image (AUC = 0.765 vs. 0.875). On the other hand, it can be seen that the filtering routine applied to the noised images provided positive results, being able to retrieve back information that was lost when high noise levels were given to the image (AUC= 0.475 vs. 0.754, for noise and filtered, respectively). It should also be noted that the results for the lowest noise value resulted in a AUC for the noised set was of 0.765 while for the filtered noised was of 0.500. This result matches with what was obtained for the MSE variation seen in section 4.1 – the MSE for the lowest noise value is higher for the filtered set than for the noise set. In this section it was proven that for a small noise value, the classifier can maintain a fair performance, and if the noise is high, the performance can be fairly maintained if the images pass through a median filtering routine before feature extraction.

When analyzing area influence in classification, area distribution for each image set was studied. It was seen that cases were correctly classified despite its location in area distribution and independently of being healthy or cancer cases. These results gave a hint that the classifier was not influenced by the area of the image, neither by the percentage of the image that is occupied by the breast, which was corroborated by the correlation results.

An increase in breast density is usually accompanied by an increase in the difficulty of reading mammograms. For that reason, a study on how the classifier is affected by breast density was performed for BI-RADS classes 1,2 and 3. Across five different BI-RADS 1 test set, the classifier achieves a mean AUC of 0.864, closer to the one obtained for the original test set – 0.875. For test sets concerning BI-RADS 2 the results were higher (mean AUC = 0.927), which seems to contradict the idea that the classifier is affected by an increase in breast density. Finally, for test sets concerning BI-RADS 3, the mean AUC obtained was of 0.906. Given these results, it can be inferred that the developed classifier was not affected by breast density, since maintained its good performance across different density classes.

%PD was evaluated as predictor of BC. To do that, new classifiers were train and tested with the same selected texture features, plus the %PD metric. The AUC values obtained were similar or even lower than the ones obtained with the original test set and given that, it was concluded that, when texture features are present, %PD does not add much relevant information. This finding matches with some studies analyzed in section 2.

A user interface application that allows a classification of the images into cancer/healthy classes was developed and it is ready to be used without the need of any additional software.

Concerning the results obtained in this part of the work, it is important to state that the metric chosen to evaluate classifier performance was AUC because not only because of its robustness, as it allowed a fair comparison between the work developed here and most of the results obtained in the literature review. It is also important to keep in mind several aspects: 1) the differentiation done here did not include only healthy tissue and diseased tissue; while the cancer images only contained malignant lesions, the healthy images contained both tissue without any lesion and images with benign findings, that did not evolve to malignancy; 2) opposite to what happens in different studies, the lesions were not of just one type, with both the benign and malignant images containing calcifications and masses with various aspects; 3) contrary to what happens in the researched papers, the texture analysis was not done only in a region that contains the lesion, but rather in the entire breast. While the presence of tissue without lesion could make the classification process easier than in the studies analyzed, the adding of benign lesions to the healthy set and the use of the entire breast instead of only a region that is known to contain the lesion, evens the balance in terms of classification difficulty. The AUC obtained here was of 0.875, a value closer or even higher than the ones obtained by the studies presented in section 2.3, being smaller than the one obtained by a research that uses previously defined and highly optimized Decision Tree classifier (AUC=0.995).

Given these considerations, it can be concluded that the algorithm developed not only has a performance close to the state-of-the art classifications for this type of task, as it is not biased by the area of each image, or by the density of the breast, achieving the proposed goals.

Finally, in terms of risk analysis, the fact that the data that allowed this analysis was so small did not allow the application of ML algorithms. However, the feature variation across years that was observed allowed to unveil what seems to be a trend in cancer patients for specific features. The fact that this trend was observed in patients that were healthy and that later were found to have a breast lesion – or to have underwent a surgical intervention involving breast parenchyma gives confidence in the use of these trends to evaluate cancer risk. However, as it is shown by the same results, the trend in feature value indicates parenchyma changes that can be due to other reasons, as the case of the patient that had a surgical procedure. For that reason, this type of applications – both for risk analysis and cancer detection – should be use cautiously, with a critical look, once they are applications to aid the diagnosis and not to substitute the physicians, in the diagnostic making process.

In terms of future work, the focus should really be in the use of machine or deep learning methodologies – with a pipeline that allows for the use of different mammogram views - in the

prediction of BC risk that was initiated in this work. The fact that there are trends in cancer patients that can be seen without the use of ML/DL algorithms and adding to that the use of the entire breast and not only a specific ROI, should motivate the use of these methodologies for risk prediction.

The development of Computer Vision algorithms that can predict BC risk can play a major role in early diagnosis of the disease, allowing for cancer to be treated in an early stage, where there is a greater chance of success. Besides that, allowing to assess future risk can aid clinicians to customize, in a patient-by-patient fashion, preventive treatment options, screening periodicity; and to manage patients expectations. These "to be developed" algorithms , along with the work developed in this thesis, provide many benefits to medical staff, BC patients, and scientists in this area. The constructed user interface can be used for medical training, to help inexperienced clinicians gaining confidence in their diagnosis, and can also serve as a ground truth checker for future works in this area of research.

As it was perceived in section 1, decreasing trends in BC mortality are not as noticeable in Black women, when compared to Caucasian women. The discrimination felt by this ethnicity across centuries (and that, although more hidden, persists today) has its impacts in the medical health field. Struggling with financial difficulties, the access to normal healthcare plans is very difficult, which ultimately leads black women to attend hospitals with less conditions, less medical staff, that it is also less trained, and with lower quantity and quality of screening machinery. The classifier developed here, through the use of the designed user interface, can have a positive impact in these less developed institutions, aiding less trained clinicians, and allowing faster mammogram analysis, which is very important in short-staffed hospitals. Therefore, besides the work done in this thesis, it is important that future work done in the area ensures ethnicity balance in training data so that it contributes to diminishing social and ethnic asymmetries that still exist in what concerns to a global basic human right – health.

6. References

[1] American Cancer Society. Breast Cancer Facts & Figures 2019-2020. Atlanta: American Cancer Society, Inc. 2019.

[2] P. Anand *et al.*, "Cancer is a preventable disease that requires major lifestyle changes," *Pharm Res,* vol. 25, no. 9, pp. 2097-2116, 2008.

[3] G. D. M. Hammer, Stephen J., "Breast Carcinoma," in *Pathophyisiology of The Disease,* 8th ed.: McGraw-Hill Education, 2019, pp. 267-273.

[4] C. E. DeSantis *et al.*, "Breast cancer statistics, 2019," *CA Cancer J Clin,* vol. 69, no. 6, pp. 438-451, 2019.

[5] National Health Service. (October 23, 2020). *Overview: Breast Cancer Screening*. Available: https://www.nhs.uk/conditions/breast-cancer-screening/

[6] American Cancer Society. (October 23, 2020). *American Cancer Society Guidelines for the Early Detection of Cancer*. Available: https://www.cancer.org/healthy/find-cancer-early/cancer-screening-guidelines/american-cancer-society-guidelines-for-the-early-detection-of-cancer.html

[7] N. Breen, J. F. Gentleman, and J. S. Schiller, "Update on mammography trends: comparisons of rates in 2000, 2005, and 2008," *Cancer*, vol. 117, no. 10, pp. 2209-2218, 2011.

[8] R. M. Pfeiffer, Y. Webb-Vargas, W. Wheeler, M. H. Gail, and P. Biomarkers, "Proportion of US trends in breast cancer incidence attributable to long-term changes in risk factor distributions,", *Cancer Epidemiol Biomarkers* , vol. 27, no. 10, pp. 1214-1222, 2018.

[9] D. A. Berry *et al.*, "Effect of screening and adjuvant therapy on mortality from breast cancer,", N Engl J Med, vol. 353, no. 17, pp. 1784-1792, 2005.

[10] R. V. Drake, A. Wayne, Mitchell, Adam, "Breast," in *Gray's Anatomy for Students,* 4th ed.: Elsevier, 2019, pp. 140-143.

[11] S. K. Saladin, "The Breasts and Mammary Glands," in *Human Anatomy,* McGraw-Hill Eduation, Ed., 2008, pp. 758-759.

[12] F. H. Netter, "Mammary Gland," in *Atlas of Human Anatomy,* Elsevier, 2018, p. 188.

[13] S. Pandya and R. G. Moore, "Breast development and anatomy," *Clin Obstet Gynecol,* vol. 54, no. 1, pp. 91-95, 2011.

[14] R. J. Santen and R. Mansel, "Benign breast disorders," *N Engl J Med,* vol. 353, no. 3, pp. 275-285, 2005.

[15] S. E. Singletary, "Rating the risk factors for breast cancer," *Ann Surg,* vol. 237, no. 4, pp. 474-482, 2003.

[16] S. Tretli, "Height and weight in relation to breast cancer morbidity and mortality. A prospective study of 570,000 women in Norway," *Int J Cancer,* vol. 44, no. 1, pp. 23-30, 1989.

[17] M. C. Pike, M. D. Krailo, B. E. Henderson, J. T. Casagrande, and D. G. Hoel, "'Hormonal' risk factors, 'breast tissue age' and the age-incidence of breast cancer," *Nature,* vol. 303, no. 5920, pp. 767-770, 1983.

[18] M. E. Barnard, C. E. Boeke, and R. M. Tamimi, "Established breast cancer risk factors and risk of intrinsic tumor subtypes," *Biochim Biophys Acta,* vol. 1856, no. 1, pp. 73-85, 2015.

[19] H. Li *et al.*, "Computerized Analysis of Mammographic Parenchymal Patterns on a Large Clinical Dataset of Full-Field Digital Mammograms: Robustness Study with Two High-Risk Datasets," (in English), *Journal of Digital Imaging,* vol. 25, no. 5, pp. 591-598, 2012.

[20] J. N. Wolfe, "Breast Patterns as an Index of Risk for Developing Breast Cancer," *American Journal of Roentgenology,* vol. 126, no. 6, pp. 1130-1139, 1976.

[21] J. N. Wolfe, "The Prominent Duct Patter as Indicator of Cancer Risk," *Oncology,* vol. 23, no.2, pp. 140-158, 1969.

[22] J. N. Wolfe, "Risk for Breast Cancer Development Determined by Mammographic Parenchymal Pattern" *Cancer,* vol. 37, pp. 2486-2492, 1976.

[23] J. N. Wolfe, "A Study of Breast Parenchyma by Mammography In the Normal Woman and Those with Benign and Malignant Disease," *Radiology,* vol. 89, no. 2, pp. 201-205, 1967.

[24] N. F. Boyd, L. J. Martin, M. J. Yaffe, and S. Minkin, "Mammographic density and breast cancer risk: current understanding and future prospects," *Breast Cancer Res,* vol. 13, no. 6, p. 223, 2011.

[25] R. W. Jakes, S. W. Duffy, F. C. Ng, F. Gao, and E. H. Ng, "Mammographic parenchymal patterns and risk of breast cancer at and after a prevalence screen in Singaporean women," *Int J Epidemiol,* vol. 29, no. 1, pp. 11-19, 2000.

[26] Z. Huo, M. L. Giger, D. E. Wolverton, W. Zhong, S. Cumming, and O. I. Olopade, "Computerized analysis of mammographic parenchymal patterns for breast cancer risk assessment: feature selection," *Med Phys,* vol. 27, no. 1, pp. 4-12, 2000.

[27] B. L. Niell, P. E. Freer, R. J. Weinfurtner, E. K. Arleo, and J. S. Drukteinis, "Screening for Breast Cancer," *Radiol Clin North Am,* vol. 55, no. 6, pp. 1145-1162, 2017.

[28] C. Coleman, "Early Detection and Screening for Breast Cancer," (in English), *Seminars in Oncology Nursing,* vol. 33, no. 2, pp. 141-155, 2017.

[29] M. Loberg, M. L. Lousdal, M. Bretthauer, and M. Kalager, "Benefits and harms of mammography screening," *Breast Cancer Res,* vol. 17, no. 1, p. 63, 2015.

[30] C. D. Lehman, E. White, S. Peacock, M. J. Drucker, and N. Urban, "Effect of age and breast density on screening mammograms with false-positive findings," *AJR Am J Roentgenol,* vol. 173, no. 6, pp. 1651-1655,1999.

[31] N. Biller-Andorno and P. Juni, "Abolishing mammography screening programs? A view from the Swiss Medical Board," *N Engl J Med,* vol. 370, no. 21, pp. 1965-1967, 2014.

[32] J. Geisel, M. Raghu, and R. Hooley, "The Role of Ultrasound in Breast Cancer Screening: The Case for and Against Ultrasound," *Semin Ultrasound CT MR,* vol. 39, no. 1, pp. 25-34, 2018.

[33] A. M. Rocha Garcia and D. Mera Fernandez, "Breast tomosynthesis: state of the art," *Radiologia,* vol. 61, no. 4, pp. 274-285, 2019.

[34] M. Morrow, J. Waters, and E. Morris, "MRI for breast cancer screening, diagnosis, and treatment," *Lancet,* vol. 378, no. 9805, pp. 1804-1811, 2011.

[35] D. G. Evans *et al.*, "MRI breast screening in high-risk women: cancer detection and survival analysis," *Breast Cancer Res Treat,* vol. 145, no. 3, pp. 663-672, 2014.

[36] D. L. Monticciolo, M. S. Newell, L. Moy, B. Niell, B. Monsees, and E. A. Sickles, "Breast Cancer Screening in Women at Higher-Than-Average Risk: Recommendations From the ACR," *J Am Coll Radiol,* vol. 15, no. 3 Pt A, pp. 408-414, 2018.

[37] U. Fischer and F. Baum, "Interventional Breast Imaging," 1st ed: Thieme, 2010, pp. 43-45.

[38] D. Ikeda, K. K. Miyake, "Mammographic Analysis of Breast Calcifications," in *Breast Imaging*: ELSEVIER, 2017, pp 75-121

[39] D. Ikeda, K. K. Miyake, "Mammographic and Ultrasound Analysis of Breast Masses," in *Breast Imaging,* 3rd ed: ELSEVIER, 2017, pp 122-170

[40] National Institute of Standards and Technology. (January 16 2021). *Engineering Statistics Handbook*. Available: https://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm

[41] R. M. S. Haralick, K. Dinstein, I. , "Textural Features for Image Classification," *IEEE Transactions on Systems, Man, and Cybernetics,* vol. SMC-E, no. 6, pp. 610-621, 1973.

[42] M. M. Galloway, "Texture analysis using gray level run lengths," *Computer Graphics and Image Processing,* vol. 4, no. 2, pp. 172-179, 1975.

[43] H. Li *et al.*, "Computerized analysis of mammographic parenchymal patterns for assessing breast cancer risk: effect of ROI size and location," *Med Phys,* vol. 31, no. 3, pp. 549-555, 2004.

[44] M. Tan, S. Mariapun, C. H. Yip, K. H. Ng, and S. H. Teo, "A novel method of determining breast cancer risk using parenchymal textural analysis of mammography images on an Asian cohort," *Phys Med Biol,* vol. 64, no. 3, p. 035016, 2019.

[45] M. Tan, J. Pu, S. Cheng, H. Liu, and B. Zheng, "Assessment of a Four-View Mammographic Image Feature Based Fusion Model to Predict Near-Term Breast Cancer Risk," *Ann Biomed Eng,* vol. 43, no. 10, pp. 2416-2428, 2015.

[46] H. Li, M. L. Giger, O. I. Olopade, A. Margolis, L. Lan, and M. R. Chinander, "Computerized texture analysis of mammographic parenchymal patterns of digitized mammograms," *Acad Radiol,* vol. 12, no. 7, pp. 863-873, 2005.

[47] S. Dhahbi, W. Barhoumi, J. Kurek, B. Swiderski, M. Kruk, and E. Zagrouba, "False-positive reduction in computer-aided mass detection using mammographic texture analysis and classification," (in English), *Computer Methods and Programs in Biomedicine,* vol. 160, pp. 75-83, 2018.

[48] M. K. Amadasun, R., "Textural features corresponding to textural properties," *IEEE Transactions on Systems, Man, and Cybernetics,* vol. 19, no. 5, pp. 1264-1274, 1989.

[49] J. N. Alzubi, A.Kumar, A., "Machine Learning from Theory to Algorithms: An Overview," presented at the Second National Conference on Computational Intelligence, India, 2018.

[50] E. I. Naqa and R. M. Li, J. M. , "What Is Machine Learning?," in *Machine Learning in Radiation Oncology: Theory and Applications*: Springer International Publishing, 2015 pp. 3-9.

[51] R. C. Deo, "Machine Learning in Medicine," *Circulation,* vol. 132, no. 20, pp. 1920-1930, 2015.

[52] C. J. M. Ang, A., H. Haron, and H. Hamed, "Supervised, Unsupervised, and Semi-Supervised Feature Selection: A Review on Gene Selection," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 13, no. 5, pp. 971-979, 2016.

[53] T. K. Fukushima, Y. Kamei, S. McIntosh, K. Yamashita, N. Ubayashi"An empirical study of just-in-time defect prediction using cross-project models," in *11th Working Conference on Mining Software Repositories*, 2014, pp. 172-181.

[54] J. Mendes and N. Matela, "Breast Cancer Risk Assessment: A Review on Mammography-Based Approaches", *Journal of Imaging*,vol. 7, no. 6, p. 98, 2021.

[55] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern recognition,* vol. 38, no. 12, pp. 2270-2285, 2005.

[56] M. J. I. Goyal, "Morphological image processing,", IJCST, vol. 2, no. 4, 2011.

[57] A. M. Raid, W. M. Khedr, M. A. El-Dosuky, M. Aoud "Image restoration based on morphological operations," *International Journal of Computer Science, Engineering and Information Technology,* vol. 4, no. 3, pp. 9-21, 2014.

[58] A. M. Hambal, Z. Pei, F. L. Ishabailu, "Image noise reduction and filtering techniques," , *International Journal of Science and Research*, vol. 6, no. 3, pp. 2033-2038, 2017.

[59] M. Abdel-Basset, A. E. Fakhry, I. El-Henawy, T. Qiu, and A. K. Sangaiah, "Feature and intensity based medical image registration using particle swarm optimization,", *Journal of Medical Systems* vol. 41, no. 12, pp. 1-15, 2017.

[60] L. G. Brown, "A survey of image registration techniques,", *ACM Computing Surveys,* vol. 24, no. 4, pp. 325-376, 1992.

[61] S. van Engeland, P. Snoeren, J. Hendriks, and N. Karssemeijer, "A comparison of methods for mammogram registration,", *IEEE Transactions on Medical Imaging,* vol. 22, no. 11, pp. 1436-1444, 2003.

[62] S. Famouri, L. Morra, and F. Lamberti, "A Deep Learning Approach for Efficient Registration of Dual View Mammography," in *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, 2020, pp. 162-172: Springer.

[63] Y. Guo, J. Suri, and R. Sivaramakrishna, "Image registration for breast imaging: a review," in *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, Shanghai, pp. 3379-3382: IEEE.

[64] L. Savita, T. Rupali, S. Almas, and D. D. Prapti, "Detection and Classification of Breast Mass Using Support Vector Machine,", *IOSR Journal of Computer Engineering (IOSR-JCE)*, The International Conference on Recent Advancements in Computing, Taiwan, pp. 1-6, 2017.

[65] A. A. Kayode, N. O. Akande, A. A. Adegun, and M. O. Adebiyi, "An automated mammogram classification system using modified support vector machine,", *Medical Devices,* vol.12, pp. 275, 2019.

[66] H. Li, K. R. Mendel, L. Lan, D. Sheth, and M. L. Giger, "Digital Mammography in Breast Cancer: Additive Value of Radiomics of Breast Parenchyma," *Radiology,* vol. 291, no. 1, pp. 15-20, 2019.

[67] A. K. Mohanty, M. R. Senapati, S. Beberta, S. K. Lenka, "Texture-based features for classification of mammograms using decision tree,", *Neural Computing and Applications,* vol. 23, no. 3, pp. 1011-1017, 2013.

[68] J. Diz, G. Marreiros, and A. Freitas, "Using data mining techniques to support breast cancer diagnosis," in *New Contributions in Information Systems and Technologies*: Springer, 2015, pp. 689-700.

[69] Y. Zheng *et al.*, "Parenchymal texture analysis in digital mammography: A fully automated pipeline for breast cancer risk assessment," *Med Phys,* vol. 42, no. 7, pp. 4149-60, 2015.

[70] M. Tan, B. Zheng, P. Ramalingam, and D. Gur, "Prediction of Near-term Breast Cancer Risk Based on Bilateral Mammographic Feature Asymmetry," (in English), *Academic Radiology,* vol. 20, no. 12, pp. 1542-1550, 2013.

[71] Z. Gandomkar *et al.*, *Breast cancer risk prediction in Chinese women based on mammographic texture and a comprehensive set of epidemiologic factors* (Fifteenth International Workshop on Breast Imaging, Belgium). SPIE, 2020.

[72] (May, 27th. 2021). *CBIS-DDSM*. Available: https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM#225166295e40bd1f79d64f04b40cac57ceca9272

[73] T. Barbu, "Variational Image Denoising Approach with Diffusion Porous Media Flow," *Abstract and Applied Analysis,* vol. 2013, p. 8 pages, 2013.

[74] R. W. Gonzalez, R, "Noise Models," in *Digital Image Processing,*4th ed.: Pearson, 2018.

[75] S. J. Lim, "Median Filtering " in *Two-Dimensional Signal and Image Processing,* Prentice Hall PTR, 1989, pp. 469-476.

[76] T. Samajdar and M. I. Quraishi, "Analysis and evaluation of image quality metrics," in *Information Systems Design and Intelligent Applications*: Springer, 2015, pp. 369-378.

[77] O. Faragallah *et al.*, "A Comprehensive Survey Analysis for Present Solutions of Medical Image Fusion and Future Directions," *IEEE Access*., vol. 9, pp. Faragallah, O.S., Elhoseny, H.M., El-Shafai, W., El-Rahman, W.A., El-sayed, H.S., El-Rabaie, E.M., El-Samie, F.E., & Geweid, G.G. (2021). A Comprehensive Survey Analysis for Present Solutions of Medical Image Fusion and Future Directions. IEEE Access, 9, 11358-11371. ,2020.

[78] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing* ,vol. 13, no. 4, pp. 600-612, 2004.

[79] *Orange Data Mining*(2021, July, 13th). Available: https://orangedatamining.com/

[80] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no.1, pp. 159-174, 1977.

[81] MathWorks®. (2021, July, 13th). *Learn optimal hyperplanes as decision boundaries*. Available: https://www.mathworks.com/discovery/support-vector-machine.html

[82] L. Rokach and O. Maimon, "Classification trees," in *Data mining and knowledge discovery handbook*, 1st ed: Springer, 2009, pp. 149-174.

[83] A. Subasi and E. Ercelebi "Classification of EEG signals using neural network and logistic regression,", *Computer methods and programs in biomedicine,* vol. 78, no. 2, pp. 87-99, 2005.

[84] A..Tharwat, "Linear vs. quadratic discriminant analysis classifier: a tutorial,", I*nternational Journal of Applied Pattern Recognition,* vol. 3, no. 2, pp. 145-180, 2016.

[85] A. Bull, "Convergence rates of efficient global optimization algorithms,", *Journal of Machine Learning Research,* vol. 12, no. 10, pp. 2879-2904, 2011.

[86] Madalena Simões, "Automatic Breast Density Classification on Tomosynthesis Images," Master of Science Biomedical Engineering, Faculty of Sciences and Technology, NOVA University Lisbon, 2020.

[87] S. Bubeck, "Convex optimization: Algorithms and complexity,", *arXiv preprint arXiv:1405.4980.*, 2014.