

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO



Pre-CADs in Breast Cancer

Joana Cristina Lopes da Fonseca

Mestrado Integrado em Engenharia Eletrotécnica e de Computadores

Supervisor: Jaime S. Cardoso (PhD)

Co-Supervisor: Inês Domingues (MSc)

July 25, 2013

A Dissertação intitulada
“Pre-CADs in Breast Cancer”

foi aprovada em provas realizadas em 15 Julho 2013

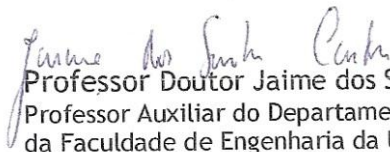
o júri



Presidente Professor Doutor Luís António Pereira de Meneses Corte-Real
Professor Associado do Departamento de Engenharia Eletrotécnica e de
Computadores da Faculdade de Engenharia da Universidade do Porto



Professor Doutor António José Ribeiro Neves
Professor Auxiliar do Departamento de Eletrónica, Telecomunicações e Informática
da Universidade de Aveiro

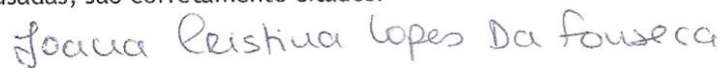


Professor Doutor Jaime dos Santos Cardoso
Professor Auxiliar do Departamento de Engenharia Eletrotécnica e de Computadores
da Faculdade de Engenharia da Universidade do Porto



Mestre Inês Domingues
Investigadora do INESC-Porto

O autor declara que a presente dissertação (ou relatório de projeto) é da sua exclusiva autoria e foi escrita sem qualquer apoio externo não explicitamente autorizado. Os resultados, ideias, parágrafos, ou outros extratos tomados de ou inspirados em trabalhos de outros autores, e demais referências bibliográficas usadas, são corretamente citados.



Autor - Joana Cristina Lopes da Fonseca

Resumo

Neste estudo é apresentado um sistema de pre-CAD, que visa ajudar os radiologistas na análise do elevado número de mamogramas que têm que avaliar a cada dia, ajudando a prevenir o elevado número de classificações erradas que podem ocorrer, devido à tarefa repetitiva a que são submetidos.

O principal objetivo deste trabalho é automatizar a classificação de mamogramas não normais. Como um sistema de pre-CAD, não falhar a classificação de nenhum mamograma não normal é o ponto mais importante. Os mamogramas cujo sistema não tem a certeza da sua classificação, serão vistos por pessoal médico. Este sistema visa ajudar o pessoal médico com aqueles mamogramas de difícil classificação, fornecendo uma previsão sólida baseada na análise de uma base de dados fidedigna.

O método proposto consiste em extrair características dos mamogramas, que foram previamente classificados de acordo com a densidade da mama para posteriormente serem classificados em normal ou não normal, para cada um dos tipos de densidade da mama.

Uma base de dados composta por 203 mamogramas foi usada. Do número total de mamogramas, 153 foram considerados normais e 50 foram considerados não normais.

Os 203 mamogramas foram classificados por especialistas, de acordo com a densidade da mama, em mamogramas densos e mamogramas adiposos. Três grupos de mamogramas foram compostos através desta classificação de densidade: um grupo composto de mamogramas densos, outro grupo composto de mamogramas adiposos e um último grupo contendo a totalidade de mamogramas usadas neste trabalho.

De cada um dos três grupos de mamogramas, dois tipos de características foram extraídos: características *Grey Level Co-occurrence Matrix* e características *Local Binary Pattern*.

Para a classificação, uma etapa crucial neste trabalho, foram usados três classificadores, de forma a ser possível estudar a *performance* de cada um dos classificadores, quando combinados com diferentes tipos de características. Os classificadores usados foram: *K-nearest neighbour*, *Support Vector Machines* e *Random Forests*.

A utilização de três diferentes cenários neste trabalho permitiu estudar não só a *performance* dos classificadores, como também o efeito da classificação prévia de acordo com a densidade e o efeito da extração das diferentes features.

Abstract

In this study we present a pre-CAD system that aims to help the radiologists in the analysis of the high number of mammograms that they have to evaluate each day, helping to prevent the increased number of misclassification that could happen, due to the repetitive task to which they are submitted.

The main objective of this work is to automate the classification of abnormal mammograms. As a pre-CAD system, not to miss any malignant mammogram is the most important point. The mammograms that the system was not sure, will be classified again by the medical staff. This system aims to help the medical staff in the mammograms that are difficult to analyse, providing an solid opinion based in the analysis of a trusted database.

The method consists in extracting features from mammograms, previously classified by experts according to the breast density and then classify them into normal or abnormal mammograms for each of the tissue density types.

A databased composed of 203 mammograms was used. Of this total number of mammograms, 153 were considered normal mammograms and 50 were considered abnormal mammograms.

The 203 mammograms were classified by experts, according to the density of the breast, into dense and fatty, and three sets of mammograms were composed: a set composed by dense mammograms, another set composed by fatty mammograms and a set composed by the total number of mammograms.

From each one of the sets, two types of features were extracted: Gray Level Co-occurrence Matrix features and Local Binary Pattern features.

For the classification task, a crucial task in this work, three classifiers were used, in order to study the performance of each one of them. It was used the K-nearest neighbour, the Support Vector Machines and the Random Forests classifiers.

The use, in this work, of three different scenarios allowed to study not only the performance of the classifiers, as well as the effect of the previous classification according to the density and the effect of the extraction of the different features.

Acknowledgements

The research, development and writing of a dissertation is, for sure, an important step on the journey of becoming an Engineer. Although the name that appears in the dissertation is mine, i would not be able to accomplish this work by myself.

I would like to thank Prof. Dr. Jaime S. Cardoso, for all the help, guidance, patience and proposed solutions provided during the realization of this Master's Dissertation.

I also would like to thank the VCMi group in INESC-TEC, for all the talks that gave me some more knowledge in other fields, as well as, to INESC-TEC and the UTM unit for providing me a workplace during this 5 months.

To Inês Domingues for the help in the beginning of the research for this Master's Dissertation, introducing me the theme and the problematic of Breast Cancer, as well as for the corrections and suggestions in the elaboration of this document.

I would like to thank my family, specially my parents, for all the patience, support, strength, pampering and for never give up on me during all this years.

To my fairy-friends for always be there for me and for the support they gave me. For listening me for hours, even when i was down with the results and the challenge that writing a dissertation is.

To Rita, Inês, João and Mylène for the relaxing moments, the tours and the dinners to take a break from the months of hard work.

To Marta, Diogo, Vasco and Fred for being, even when they are not, always by my side, supporting and not judging me.

Finally, i would like to thank all true friends for the patience, encouragement to work, jokes when i was more discouraged, encouragement to rest when it was needed, the corrections in the writing and for, in the end, being my friends.

This research was conducted under the close sponsorship of the project BCS - Módulos Automáticos de Auxílio ao Rastreamento e Diagnóstico do Cancro da Mama Integrados em Sistemas PACS (nº 33928), funded by the European Regional Development Fund (ERDF) through the COMPETE - Operational Programme for Competitiveness Factors (POFC), project QREN-R&D in co-promotion sponsored by Portuguese Innovation Agency.

Joana Lopes da Fonseca

“It’s supposed to be automatic, but actually you have to push this button.”

John Brunner

Contents

1	Introduction	1
1.1	Context	1
1.2	Motivation	1
1.3	Objectives	2
1.4	Contributions	2
1.5	Structure	3
2	Background	5
2.1	Breast	5
2.2	Mammography	6
2.2.1	Screening Mammography	7
2.2.2	Diagnostic Mammography	7
2.3	Missed Cancers and False Positives	7
2.3.1	Missed Cancers	8
2.3.2	False Positives	8
2.4	Abnormalities	8
2.4.1	Asymmetric Density	8
2.4.2	Architectural Distortion	9
2.4.3	Calcifications	10
2.4.4	Masses	11
2.5	BI-RADS	13
3	State-of-the-art	17
3.1	Features	17
3.1.1	Curvilinear Features	17
3.1.2	Local Binary Pattern Features	20
3.1.3	Grey Level Co-occurrence Matrix Features	22
3.1.4	Multiresolution Features	26
3.1.5	Gabor Features	27
3.2	Classification Methods	28
3.2.1	K-nearest Neighbour	28
3.2.2	Decision Trees	28
3.2.3	Random Forests	30
3.2.4	Artificial Neural Networks	31
3.2.5	Support Vector Machines	32
3.2.6	No Classification Method	33
3.3	Density Classification	34
3.4	Summary	36

4	Methodology	37
4.1	Database	38
4.2	Pre-Processing of the Mammograms	39
4.3	Features Extraction	41
4.3.1	GLCM Features	41
4.3.2	LBP Features	43
4.4	Classification	46
4.4.1	K-Nearest Neighbour	46
4.4.2	Support Vector Machines	49
4.4.3	Random Forests	52
4.5	Evaluation	54
5	Results	57
5.1	Extraction of GLCM and LBP features with previous density classification	57
5.1.1	K-Nearest Neighbour Classifier	57
5.1.2	Support Vector Machines Classifier	59
5.1.3	Random Forests Classifier	61
5.1.4	Conclusion Scenario 1	63
5.2	Extraction of features with previous density classification	64
5.2.1	Dense	64
5.2.2	Fatty	65
5.2.3	Conclusion Scenario 2	66
5.3	Extraction of GLCM and LBP Features without previous density classification	67
5.3.1	Conclusion Scenario 3	70
5.4	Conclusions	70
6	Conclusions and Future Work	73
6.1	Conclusions	73
6.2	Future Work	75
	References	77

List of Figures

2.1	Breast Anatomy and Mammography (from [1])	6
2.2	Normal Mammography (from [1])	6
2.3	Cranial-Caudal View and Mediolateral-Oblique View (from [2])	7
2.4	Lateromedial View and Mediolateral View (from [2])	7
2.5	Mammography with Asymmetric Density (from [3])	8
2.6	Mammography with Architectural Distortion (from [4])	9
2.7	Mammography with Micro-calcifications (from [4])	11
2.8	Mammography with Cluster of Micro-calcifications (from [5])	11
2.9	Mammography with a mass (from [4])	12
2.10	Shape mass descriptors and Margin mass descriptors (from [6])	12
2.11	Bi-RADS Categories Assessments (from [7])	13
4.1	General Block Diagram of the Proposed Methodology: A, B - GLCM; C - SVM, Random Forests, K-nearest neighbour	37
4.2	Distribution of Dense and Fatty Mammograms according to BI-RADS class	38
4.3	Determination of most high, low and extreme points	41
4.4	Scheme directions calculations GLCM	42
4.5	Scheme calculation GLCM (8 levels) for 0° or $[0 \ d]$	43
4.6	Scheme of basic LBP	44
4.7	Scheme of Multi-Resolution approach to grey-scale LBP	44
4.8	Example of KNN	48
5.1	Results of the KNN classifier	59
5.2	Results of the SVM classifier	61
5.3	Results of the RF classifier	62
5.4	Results of the extraction of features together (GLCM and LBP) from dense mam- mograms with KNN and SVM classifiers	65
5.5	Results of the extraction of features together (GLCM and LBP) from fatty mam- mograms with KNN and SVM classifiers	66
5.6	Results of the extraction of GLCM features without previous density classification	68
5.7	Results of the extraction of Simple LBP features without previous density classi- fication	69
5.8	Results of the extraction of Multi-Resolution LBP features ($LBP_{8,1}$) without pre- vious density classification	69
5.9	Results of the extraction of Multi-Resolution LBP features ($LBP_{16,2}$) without pre- vious density classification	70

List of Tables

2.1	Calcifications Characteristics [4] [8]	10
4.1	Distribution of Dataset according to the second scenario	39
5.1	Results of the KNN Classifier for GLCM and LBP Features - Dense Tissue	58
5.2	Results of the KNN Classifier for GLCM and LBP Features - Fatty Tissue	58
5.3	Results of the SVM Classifier for GLCM and LBP Features - Dense Tissue	60
5.4	Results of the SVM Classifier for GLCM and LBP Features - Fatty Tissue	60
5.5	Results of the RF Classifier for GLCM and LBP Features - Dense Tissue	61
5.6	Results of the RF Classifier for GLCM and LBP Features - Fatty Tissue	62
5.7	Results for Dense tissue for KNN and SVM classifiers	64
5.8	Results for Fatty tissue for KNN and SVM classifiers	65
5.9	Results SVM Classifier for GLCM and LBP Features without previous density classification	67

Abbreviations

AD	Architectural Distortion
ASD	Asymmetric Density
ACR	American College of Radiology
ANN	Artificial Neural Network
BC	Breast Cancer
BTC	Binary Tree Classifier
CAD	Computer-aided Diagnosis or Detection
CC	Cranial-caudal View
CNN	Convolution Neural Network
DICOM	Digital Imaging and Communications in Medicine
DM	Digital Mammography
FFDM	Full-field Digital Mammography
FN	False Negative
FP	False Positive
GLCM	Gray Level Co-occurrence Matrix
GLDS	Gray Level Difference Statistics
INE	Instituto Nacional de Estadística
KKT	Karush-Kuhn-Tucker
LBP	Local Binary Pattern
LM	Lateromedial
MatLab	Matrix Laboratory
MC	Missed Cancers
ML	Mediolateral View
MLO	Mediolateral-oblique
RF	Random Forests
ROI	Region of Interest
SGLD	Spatial Gray Level Dependence
TN	True Negative
TP	True Positive

Chapter 1

Introduction

This chapter presents the context in which this work was proposed, as well as the motivation that led to its elaboration.

1.1 Context

This work arises in the context of an MSc Dissertation in Electrical and Computers Engineering scheduled to take place during the second semester of the current academic year to partially fulfil the requirements for the degree of Master of Science in Electrical and Computer Engineering - Telecommunications, Electronic and Computers Major.

1.2 Motivation

Nowadays, due to the increased number of Breast Cancer diagnosis in women, radiologists have to analyse a considerable number of mammograms everyday. One of the problems of mammography analysis is that many of them are produced at low contrast, which causes the non detection of tumours by radiologists. Statistics show that 30% of breast cancers may be missed at mammography [9]. In order to decrease the rate of missed cancers, double reading has been applied. Double reading means that for each mammography two radiologists must analyse it to prevent a missed cancer. This proved to be effective - the number has decreased, but the repetitiveness of the task plus the difficulty of interpreting the "hard to classify cases", the quality and the large number of mammograms analysed each day can contribute to a still high number of missed cancers..

In the last few years, more attention has been given towards investigating automatic detection systems that can aid radiologists in detecting suspicious regions and therefore diagnosing as early as possible. These automatic systems are called Computer-aided Diagnosis or Detection (CAD) systems and are considered a "second look" in the detection of breast cancer. It is already proven that using a CAD system will improve the detection of breast cancer and, at the same time, will decrease the number of deaths per year. In the United States, since CAD systems started being used, the mortality per year has dropped between 2 and 3.2%. The field of CAD systems is still an

open field for research because the accuracy of CAD systems remains below the level that would lead to an improvement in the overall radiologists performance [10].

Since only 0.58% of the mammograms analysed by radiologists are effectively holders of breast cancer, thereby we propose the elaboration of a "pre-CAD" system that functions as "first look" in the analysis of the mammography, filtering the mammograms that will be effectively analysed by radiologists. In practice, the pre-CAD system will separate the cases that are definitely not harmful, leaving only the cases that are hard to analyse or suspicious for the radiologists to analyse. The work of radiologists will be reduced, providing an opportunity for the radiologists to enhance their performance among the most difficult cases.

1.3 Objectives

This dissertation has four main objectives:

- Automate the classification of normal mammograms
- Understand the importance of density classification
- Determine which features work better in each tissue type
- Determine which classifier works better in each tissue type

It is important to note that the main objective of this dissertation is automate the classification of normal mammograms. Having this point working, means that the radiologists can focus in the hardest-to-classify cases, saving time that may cost lives.

1.4 Contributions

The work developed represents a step towards a more automatic pre-CAD system.

It was possible to draw some conclusions regarding the functioning of the system:

- The influence of the previous density classification.
- The influence of using, in the classification phase, the extracted features separately by type.
- Which classifier is more efficient between the three used classifiers.
- Which type of features is more suitable for dense mammograms, as well as for fatty mammograms.
- Between the two types of LBP features, which accomplishes better results.
- Which version of the LBP features is, usually, more efficient.

These important steps may help, in the future, to develop a pre-CAD system more accurate, that may be implemented in the hospitals, really helping the medical staff in the difficult task that classifying a mammogram is.

During the development of this Master's Dissertation, a paper was submitted to StudECE2013, the second PhD Students Conference in Electrical and Computer Engineering. The paper was accepted to a poster's session that took part in 26th June of 2013.

1.5 Structure

This document is divided in 6 chapters. The chapter 1, where this section is inserted is the introduction and contains not only the structure of the document, but also the context and the motivation that leads to the development of this work. The chapter 2 presents the background on the Breast Cancer problematic. It is also presented a chapter including the state-of-the-art on this theme, the chapter 3. In chapter 4 it is presented the methodology adopted to the realization of this work, as well as an explanation of the implementation of the different steps in its realization. Before the conclusion chapter, this document contains the results chapter. On the results chapter, the chapter 5, are presented the results of the elaboration of this work, as well as an analysis to those results. The last chapter, the chapter 6, presents the conclusions of this work.

Chapter 2

Background

This chapter presents the Background of the problematic of Breast Cancer, in order to understand all that concerns the breast and the breast cancer. The types of mammography, the concept of missed cancers and false positives, the abnormalities that may appear in the breast and the BI-RADS classification are the point that will be explained.

Breast cancer (BC) is the leading cause of death, by cancer, among Portuguese women. According to Instituto Nacional de Estatística (INE) the number of deaths by BC is assuming an increasing perspective. From 2006 to 2010 the standardised mortality rate for BC has increased from 3.7 to 30.3 deaths per 100000 women [11]. Although women are the most affected genre by BC, approximately 1% of all BC diagnosed per year are in men [12].

One out of every eleven women will, at some time in their lives, develop BC [13]. Each year, 4500 new cases of BC are diagnosed and 1500 women will die of BC. To better understand the numbers, only today, 11 to 13 women will be diagnosed with BC and 4 will die.

If early detected and then submitted to a correct treatment, 90% of all BC diagnosed are curable, the mortality rate could decrease up to 30% [13] [14]. The only reliable method that makes possible the detection of abnormalities up to two years before they are palpable is Mammography [15].

2.1 Breast

Each mammary gland lies between the pectoralis major muscle and the skin and is typically composed of 15-20 lobes covered by a considerable amount of fatty tissue. The lobes of each mammary gland form a conical bulk with the nipple located at the vertex, which together with the fatty tissue give the breast their conical form. Each lobe has a single lactiferous duct which ends, independently of the others in the surface of the nipple. The milk is accumulated in the main duct. The breasts are supported and held in their position by a group of suspensory ligaments, the Cooper ligaments. These ligaments form a natural bra that helps hold the breast up and keeps it from sagging. The group of glandular tissue and the supporting ligaments is called dense tissue [16]. The constitution of the breast can be seen in Figure 2.1.

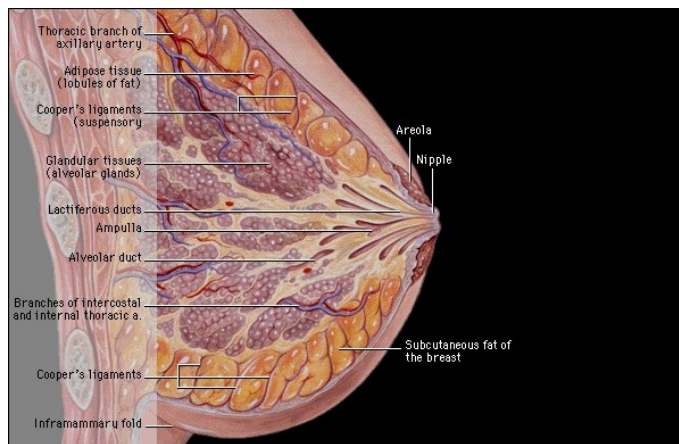


Figure 2.1: Breast Anatomy and Mammography (from [1])

2.2 Mammography

A mammography is an high resolution x-ray of the compressed breast. Mammograms require very small doses of radiation, however, comparing with other x-rays, mammograms require higher x-ray exposure because the breast is totally composed of soft tissue and has very low contrast [17].

Normal mammography image (Figure 2.2) shows a thin regular skin line with a diffuse, even, soft tissue density of the general glandular tissue and fatty structures organised in a relatively regular way by Cooper's ligaments [1].

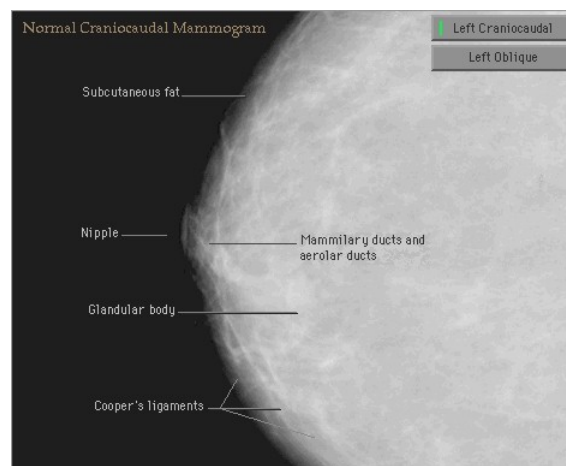


Figure 2.2: Normal Mammography (from [1])

There are two types of mammography: screening mammography and diagnostic mammography, described in the following sections.

2.2.1 Screening Mammography

Screening Mammography is characterised by being performed towards to early detection. It's performed inside a regular period and consists of two views of each breast: Cranial-Caudal (CC) and Mediolateral-Oblique (MLO) (Figure 2.3).

According to Administração Regional de Saúde - NORTE (ARS - NORTE), women between 45 and 69 should have a mammogram every two years, unless the previous examinations require another periodicity [18].

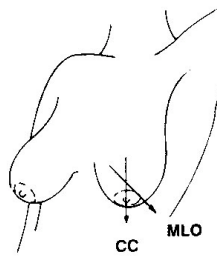


Figure 2.3: Cranial-Caudal View and Mediolateral-Oblique View (from [2])

2.2.2 Diagnostic Mammography

This type of mammography is characterised by being performed when there are suspicious abnormalities in the breast and it is necessary to make better viewing images for confirmation. Diagnostic mammography is comprised by four views: CC, MLO, Lateromedial (LM) and Mediolateral (ML) in each breast (Figure 2.3 and 2.4).

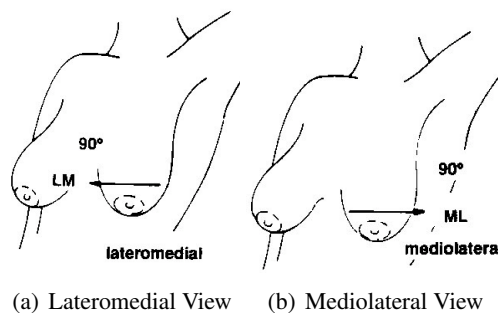


Figure 2.4: Lateromedial View and Mediolateral View (from [2])

2.3 Missed Cancers and False Positives

In this section we will discuss situations that if occurring in the analysis of a mammography, will affect the correct detection of the cancer.

2.3.1 Missed Cancers

Missed Cancers are all the cancers not detected as cancer in the mammography, in other words, the False Negatives (FN). FN occur when mammograms appear normal, but in reality there is cancer present in the breast. The main cause of FN results is high breast density and the problem associated to distinguish abnormalities in dense tissue.

2.3.2 False Positives

False positives occur when after analysis of qualified radiologists, they decide that a mammogram is abnormal but no cancer is actually present. All irregular mammograms should require additional tests to, in this way, determine if there is actually cancer present [19].

2.4 Abnormalities

There is an impressive number of abnormalities that could be detected in a mammography, but in this section only the four most recurrent ones are focused. Of these four, calcifications and masses are the ones that most commonly appear in breast cancer scenarios.

2.4.1 Asymmetric Density

Asymmetric Density (ASD) refers to a greater volume or density of breast tissue in one breast than in the corresponding area in the contralateral breast (Figure 2.5). It is a fairly vague finding in which there is no focal mass, no distorted architecture, no central density and no associated breast calcifications. It is an abnormality that appears in 3% of all mammograms. Only a few percentage of them will actually be taken to biopsy, and even a smaller percentage will truly have breast cancer [4].

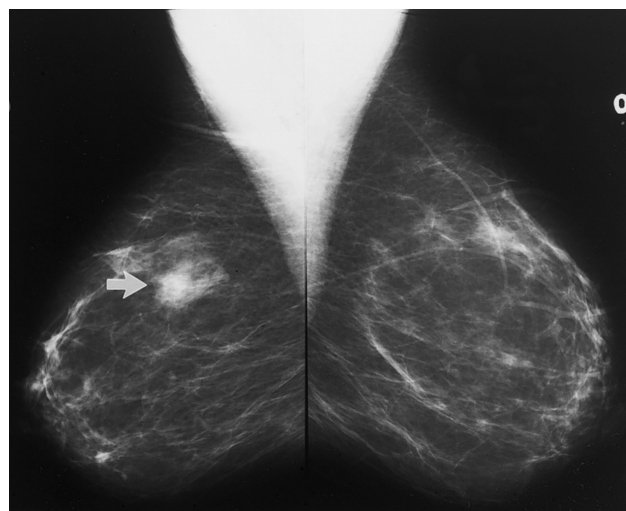


Figure 2.5: Mammography with Asymmetric Density (from [3])

Although asymmetry is often a normal finding, additional evaluation may sometimes be required. Asymmetry may be secondary to removal of tissue or to lack of development or more prominent parenchyma in one breast [3]. ASD should only be taken into account when associated with a clinically palpable breast asymmetry.

2.4.2 Architectural Distortion

Architectural Distortion (AD) is basically a distortion of the normal "random" pattern of curvilinear and fine linear radiopaque¹ structures normally observed in a mammography. There is no visible mass but the distortion often appears as a "stellate" shape or radiating spiculations that radiate from a common point (Figure 2.6). AD may be associated with BC because cancer infiltration can disrupt parenchymal architecture before there is evidence of a mass. It can be also seen in areas of prior injury or breast surgery that tend to improve or remain over time which means that comparison with previous findings is essential. Patients with AD are more likely to have positive margins² than patients with masses or calcifications and can also be said that in a mammography, tumours presented as AD are also bigger than the ones presented as other abnormalities. That fact could happen because most AD found in a mammography are due to benign causes. The number of those women which the AD would actually represent invasive breast cancer is very low [4] [3] [20].

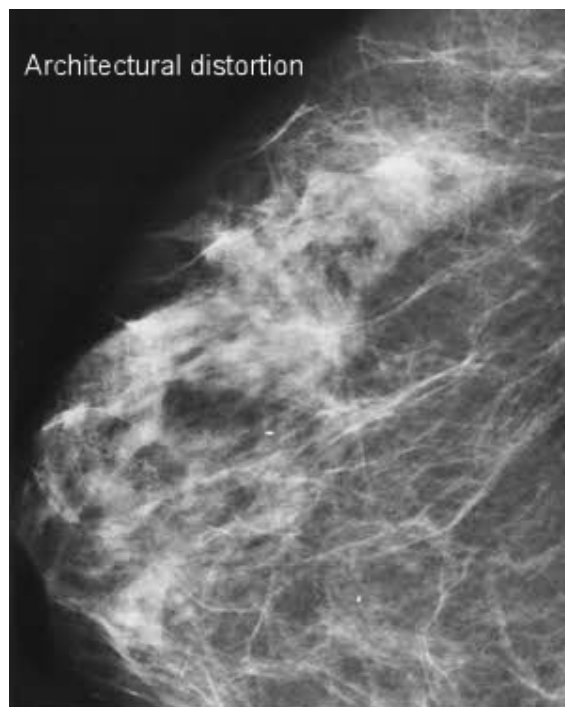


Figure 2.6: Mammography with Architectural Distortion (from [4])

¹Opaque to the radiation transmitted in the mammography.

²Cancer cells extend out to the edge of the tissue where the ink is.

2.4.3 Calcifications

Calcifications are small calcium deposits in the breast. They appear in the mammography as high density regions and are due to cell secretion and necrotic cellular debris. Calcifications are extremely common among women and could appear with or without associated lesion. The shape and distribution of calcifications provides extremely viable clues about the benign and malignancy condition of the breast lesion.

A summary with the principal characteristics of calcifications can be seen in Table 2.1.

Table 2.1: Calcifications Characteristics [4] [8]

<i>Morphology</i>	<i>Distribution</i>	<i>Size</i>	<i>Density</i>
Amorphous	Clustered	Small(<200 μ m)	Homogeneous
Branching	Linear	Big	Heterogeneous
Pleomorphic	Multiple Clusters		
Punctate	Segmental		
Round			

The conductive line in relation to the condition clues is that larger, round or oval shaped calcifications with a uniform size have a higher probability of being associated with a benign condition, while calcifications that are smaller, irregular, polymorphic and branching with a heterogeneous size and morphology are more often associated with a malignant condition.

Typically individual micro-calcifications range in size from 0.1 up to 1.0mm having an average diameter of 0.5mm. Clusters have normally 1cm² of area [15].

There are two types of calcifications: micro-calcifications and macro-calcifications, which will be explained in more detail below [15].

2.4.3.1 Micro-calcifications

Micro-calcifications are also calcium deposits.

They can appear isolated (Figure 2.7), in clusters³ (Figure 2.8) or found embedded in a mass. Clusters are important findings in a mammography because they give clues in the determination of the importance of the abnormality. 30 to 50% of non-palpable cancers are initially detected due to the presence of clusters in the mammography [4] [15].

³Clusters are a group of usually 3 or more individual micro-calcifications.

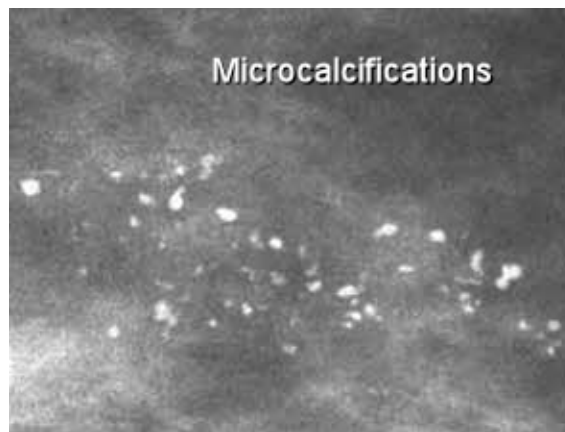


Figure 2.7: Mammography with Micro-calcifications (from [4])

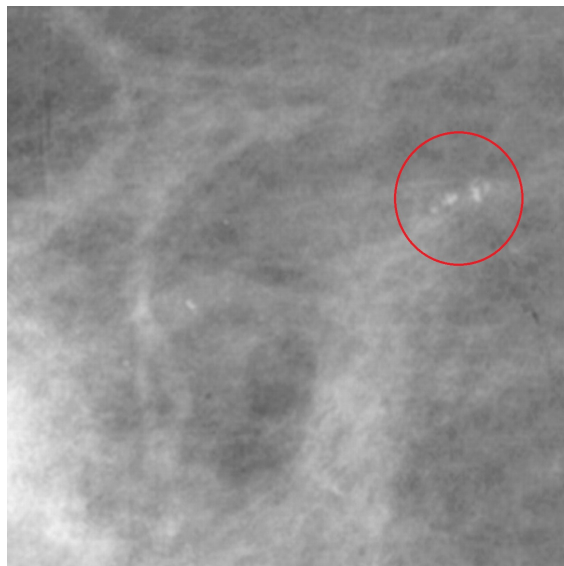


Figure 2.8: Mammography with Cluster of Micro-calcifications (from [5])

2.4.3.2 Macro-calcifications

Macro-calcifications are coarse, scattered calcium deposits and usually are not associated with malignant conditions.

2.4.4 Masses

Masses are three-dimensional lesions that when observed in at least two views of a mammography and felt in previous palpable examination, may represent a signal of BC [6] (Figure 2.9).

Masses are more difficult to detect than calcifications due to the similarity between the intensities of the normal breast tissue and the constitution of the mass. Masses are detected through a

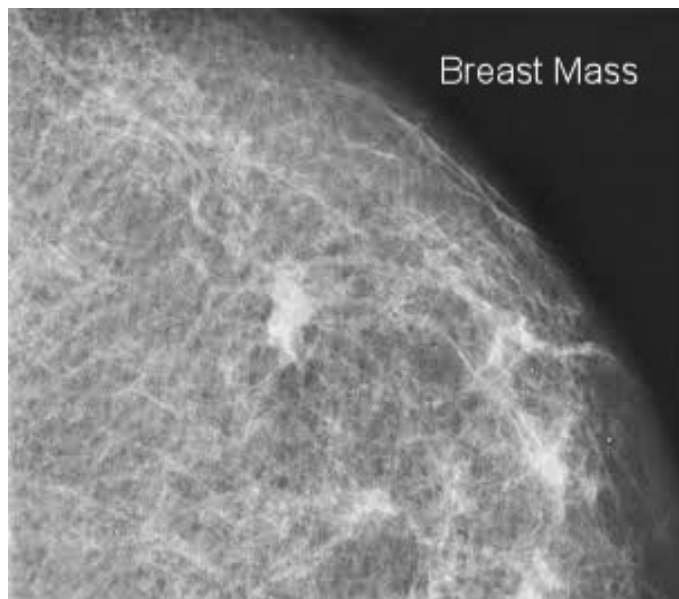


Figure 2.9: Mammography with a mass (from [4])

group of characteristics such as shape, size and density of the mass, as well as the characteristics of the margins [15]. The characteristics of the mass shape and the margins can be seen in Figure 2.10.

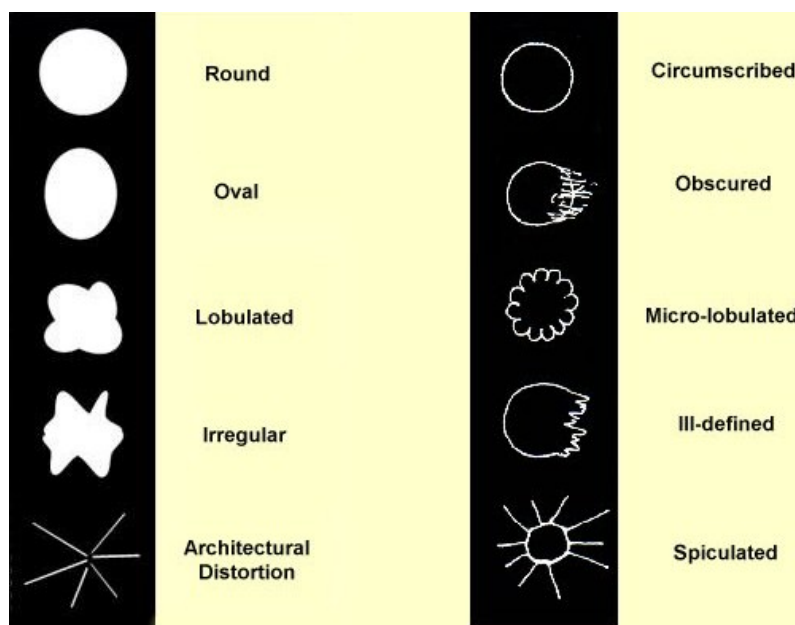


Figure 2.10: Shape mass descriptors and Margin mass descriptors (from [6])

Through the analysis of the masses characteristics, it is possible to predict whether a mass is benign or malignant. Usually most benign masses are well circumscribed, compact and with a

round or oval shape, on the other and, malignant masses usually have an irregular shape appearance and spiculated margins [15].

2.5 BI-RADS

The American College of Radiology (ACR), in a collaborative effort between various clinical organizations, such as the National Cancer Institute and The American Medical Association, among others, developed the Breast Imaging Reporting and Data System (BI-RADS) [21].

This standard was developed with the primary objective of providing an orientation line for those health professionals that are responsible for the evaluation of mammograms and diagnosis of BC. This guidance should help them regarding the action to be taken in the presence of several mammographic findings, whether they are negatives, benignant, probably benignant, suspicious or highly suspicious. It is very useful as predictor of malignancy because allows the access to the predicted values of the mammographic findings [22]. It is important to note that, if two different findings in a single mammography are detected, it should be classified according to the finding with the higher BI-RADS category.

Breast Imaging Reporting and Database System (BI-RADS®)		
Category	Assessment	Follow-up Recommendations
a. Assessment is Incomplete		
0	Need Additional Imaging Evaluation and/or Prior Mammograms for Comparison	Additional imaging and/or prior images are needed before a final assessment can be assigned
b. Assessment is Complete – Final Categories		
1	Negative	Routine annual screening mammography (for women over age 40)
2	Benign Finding(s)	Routine annual screening mammography (for women over age 40)
3	Probably Benign Finding – Initial Short-Interval Follow-Up Suggested	Initial short-term follow up (usually 6-month) examination
4	Suspicious Abnormality – Biopsy Should Be Considered Optional subdivisions:* 4A: Finding needing intervention with a low suspicion for malignancy 4B: Lesions with an intermediate suspicion of malignancy 4C: Findings of moderate concern, but not classic for malignancy	Usually requires biopsy
5	Highly Suggestive of Malignancy – Appropriate Action Should Be Taken	Requires biopsy or surgical treatment
6	Known Biopsy-Proven Malignancy – Appropriate Action Should Be Taken	Category reserved for lesions identified on imaging study with biopsy proof of malignancy prior to definitive therapy

* A subdivision may be used **in addition to** the Category 4 final assessment; MQSA does **not** allow a subdivision to replace a Category 4 final assessment. Use of subdivision is at the discretion of the facility it is not required by the FDA.

Figure 2.11: Bi-RADS Categories Assessments (from [7])

In Figure 2.11, we can see the current edition of the BI-RADS classification. On the previous edition, the 4th category did not have sub-categories. They were inserted due to the amount of

lesions that are covered by this category. There are 3 sub-categories: 4A, for cases that are slight suspicious, 4B, for cases that are moderately suspicious and 4C for the cases that have a strong suspicion. The 6th category was also inserted in the 4th edition of the Mammography Atlas. This new category covers all cases that have already been examined by biopsy, with a malignant result but, the lesion has not been completely treated [23]. Each one of the categories of BI-RADS will be further explained in the next subsections [21] [24] [25].

Category 0

Classification with category 0 should only be used in a screening mammography context and should rarely be used after a full imaging work. If a mammogram is classified with category 0, which means that the assessment is not yet complete, the patient should do additional exams or the health professional should provide comparison with previous exams. Recommendation for additional imaging evaluation includes the use of spot compression, magnification, special mammographic views or ultrasounds. Normal findings in mammograms classified with this category include symmetrical breasts without masses, architectural distortion or micro-calcifications.

Category 1

This category is, in almost everything, similar to category 0, but in this case, the assessment is complete and has negative result.

Category 2

The risk of cancer in this category is similar to category 1, 0%. In this category, the radiologist will describe the findings, although they are benign.

Category 3

The probability of malignancy of this category is less than 2% and it is not expected that the appearance of the finding be altered during the follow-up. Although the appearance does not change, the radiologist prefers to document the stability of the finding, and if it remains indeed stable during the follow-up period, it will be classified as BI-RADS 2. On the other hand, if changes are verified in the finding during that period, it will be classified as BI-RADS 4 or 5. Because of the conditions in this category, it is necessary to do additional exams in the follow-up period. This follow-up period can be 6, 12 or 24 months.

Category 4

The probability of malignancy of this category is situated between 2 and 95%. The findings classified with this category are those without classic appearance of malignancy but that have probability of being malignant.

4A

Slight index of suspicion. The follow-up of these lesions may indicate a different diagnosis, confirming the malignancy or proving the benign condition of the finding. Solid mass partially circumscribed with characteristics suggesting complex cyst or probable tumor are examples of findings that we could find in sub-category 4A.

4B

Moderate index of suspicion. Indistinct and partial circumscribed mass are examples of lesions included in this sub-category. To classify a mammogram with this sub-category an agreement between all members involved in the diagnosis is required.

4C

The findings in this category have a high index of suspicion but they do not have the classic characteristics of malignant findings. An irregular and ill-defined solid mass or a new group of pleomorphic micro-calcifications are examples to be included in this sub-category.

Category 5

The probability of malignancy of this category is >95%. The normal findings of this category are the classic findings of breast cancer. Spiculated masses, segmental or linear micro-calcifications or spiculated and irregular masses associated to pleomorphic micro-calcifications are some of these classic findings. All the other lesions with high rate of suspicion, but that don't fit the classic criteria of cancer should be classified with the category 4.

Category 6

Reserved for lesions already diagnosed in mammograms with evidence of malignancy. It is used for classification of findings of a mammogram done after neoadjuvant chemotherapy⁴ or for revisions of previous diagnosis.

⁴Chemotherapy done before the surgery.

Chapter 3

State-of-the-art

This chapter presents the state-of-the-art of the field of pre-CADs in Breast Cancer. The main characteristics will be presented and discussed.

The main point of this pre-CAD system, features extraction and the classification will be presented in terms of the state-of-the-art, to be possible to understand what has been done in this field, as well as, which tools have been used.

3.1 Features

The extraction of features, when trying to implement an automated system of identification of normal or abnormal mammograms, is one of the main steps of the methodologies in this field. There are several sets of features that could be extracted for posterior analysis of performance for both density classification or normal/abnormal classification. The descriptors of the different features have been presented in Chiracharit *et al* and it will be presented in this work in the subsequent subsections [26].

3.1.1 Curvilinear Features

Normal breast tissue could present several different forms of appearance but, unequivocally, normal breast areas are characterised by curvilinear markings that are the ductal structures of the breast tissue. A line detection algorithm is used to extract curvilinear features in the different regions [26].

Let $f(i, j)$ be the pixel grey level at spatial location (i, j) and let $L(\theta, l)$ be a string of pixels in the direction θ and of length l . $N_{L(\theta, l)}$ is the number of pixels within $L(\theta, l)$. Then the standard deviation of the pixel grey level in $L(\theta, l)$:

$$\sigma(\theta, l) = \sqrt{\frac{\sum_{(m,n) \in L(\theta, l)} (f(m, n) - \bar{f}_{L(\theta, l)})^2}{N_{L(\theta, l)} - 1}}, \quad (3.1)$$

where $\bar{f}_{L(\theta, l)}$ is the average grey level within $L(\theta, l)$

$$\bar{f}_{L(\theta,l)} = \frac{1}{N_{L(\theta,l)}} \sum_{(m,n) \in L(\theta,l)} f(m,n), \quad (3.2)$$

Let $\sigma_{i,j}(\theta, l) = \min_{(i,j) \in L(\theta,l)} \sigma(\theta, l)$ and $\sigma_{i,j}(l) = \min_{\theta} \sigma_{i,j}(\theta, l)$. The measure of surrounding pixel difference can be obtained from the standard deviation of $\sigma_{i,j}(\theta, l)$ with regard to θ

$$\sigma_{\sigma(i,j)}(l) = \sqrt{\frac{\sum_{\theta} (\sigma_{i,j}(\theta, l) - \bar{\sigma}_{i,j}(\theta, l))^2}{N_{\theta} - 1}}, \quad (3.3)$$

where N_{θ} is the total number of directions and $\bar{\sigma}_{i,j}(\theta, l)$ is the average value of $\sigma_{i,j}(\theta, l)$. Finally, each pixel (i, j) is determined to be as a line pixel or not according to the following rule for binary curvilinear lines:

$$CL_{bin}(i, j) = \begin{cases} 1(\text{line}) & \text{if } \sigma_{i,j}(l) < T_{\sigma}, \\ & \sigma_{i,j}(l) > T_{\sigma_{\sigma}}, \\ 0(\text{not}) & \text{otherwise} \end{cases} \quad (3.4)$$

where T_{σ} and $T_{\sigma_{\sigma}}$ are thresholds determined experimentally. l_i is the line pixel count and a_i is the average angle of each sub-block, where $i = 1, 2, \dots, 4,096$. $sb \times sb$ is the height and width of the sub-blocks of the image with size $im \times im$. Histograms of h bins are obtained for l_i and a_i with the relative frequency p_j^l and bin value x_j^l at bin j .

Line Pixel Count

Line Pixel Count is the total number of curvilinear pixels in the region.

Upper Right Half Line Pixel Count

Upper Right Half Line Pixel Count is the line pixel count in the upper right diagonal half of a region and it is calculated using the following relation:

$$A = \sum \sum_{i < j} CL_{bin}(i, j) \quad (3.5)$$

Lower Left Half Line Pixel Count

Lower Left Half Line Pixel Count is the line pixel count in the lower left diagonal half of a region and it is calculated using the following relation:

$$B = \sum \sum_{i \geq j} CL_{bin}(i, j) \quad (3.6)$$

Upper Left Half Line Pixel Count

Upper Left Half Line Pixel Count is the line pixel count in the upper left diagonal half of a region and it is calculated using the following relation:

$$C = \sum_{i+j < im} \sum CL_{bin}(i, j) \quad (3.7)$$

Lower Right Half Line Pixel Count

Lower Right Half Line Pixel Count is the line pixel count in the lower right diagonal half of a region and it is calculated using the following relation:

$$D = \sum_{i+j \geq im} \sum CL_{bin}(i, j) \quad (3.8)$$

Half Ratio

Half Ratio is defined as the ratio between the Upper Right and the Lower Left Half Line Pixel Count.

$$HalfRatio = \frac{A}{B} \quad (3.9)$$

Half Ratio 2

Half Ratio 2 is defined as the ratio between the Upper Left and the Lower Right Half Line Pixel Count.

$$HalfRatio2 = \frac{C}{D} \quad (3.10)$$

Angle Mean

Angle Mean is the average line angle of the curvilinear pixels in the region.

Angle Standard Deviation

Angle Standard Deviation is the standard deviation of the angle of the curvilinear pixels in the region.

Local Line Mean

Local Line Mean is the average value of the lines pixel count and it is calculated using the following relation:

$$LLM = \sum_{j=1}^h p_j^l x_j^l \quad (3.11)$$

Local Line Standard Deviation

Local Line Standard Deviation is the standard deviation value of the lines pixel count and it is

calculated using the following relation:

$$LLS = \sum_{j=1}^h p_j^l (x_j^l - LLM)^2 \quad (3.12)$$

Local Line Entropy

Local Line Entropy is the entropy value of the lines pixel count and it is calculated using the following relation:

$$LLE = \sum_{j=1}^h -p_j^l \log p_j^l \quad (3.13)$$

Local Angle Mean

Local Angle Mean is the average value of the local angles count and it is calculated using the following relation:

$$LAM = \sum_{j=1}^h p_j^a x_j^a \quad (3.14)$$

Local Angle Standard Deviation

Local Angle Standard Deviation is the standard deviation value of the local angles count and it is calculated using the following relation:

$$LAS = \sum_{j=1}^h p_j^a (x_j^a - LAM)^2 \quad (3.15)$$

Local Angle Entropy

Local Angle Entropy is the entropy value of the local angles count and it is calculated using the following relation:

$$LAE = \sum_{j=1}^h -p_j^a \log p_j^a \quad (3.16)$$

3.1.2 Local Binary Pattern Features

Local Binary Pattern (LBP) provides discriminative texture information. The main advantage of using LBP features is the fact that these features have a very low computational complexity, which makes real-time image analysis possible. Another advantage of using LBP features is their invariance against monotonic grey level changes [27]. The basic LBP operator labels the pixels of an image by thresholding the 3×3 neighbourhood of each pixel with the center pixel value and considering the result as a binary number [28].

Given a pixel in the image, the LBP code can be computed using

$$LBP_{P,R}(x,y) = \sum_{p=0}^{P-1} S(g_p - g_c)2^p \quad (3.17)$$

where,

$$S(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (3.18)$$

g_c is the grey level value of the pixel of the center and g_p is the grey level value of its neighbours. P is the total number neighbours of the central pixel. After LBP pattern of each pixel is identified, an histogram is used to represent the texture image:

$$H(K) = \sum_{i=1}^I \sum_{j=1}^J f(LBP_{P,R}(i,j),k), k \in [0, K] \quad (3.19)$$

The histogram contains h_{lbp} bins with relative frequency p_j^{lbp} of bin x_j^{lbp} at bin j .

$$f(x,y) = \begin{cases} 1, & x = y \\ 0, & otherwise \end{cases} \quad (3.20)$$

There are different sets of features based on LBP, which is the case of:

Basic LBP

In this case, the binary code is built by thresholding the neighbourhood by the value of the center. The features extracted using this approach are simply the contrast value of the binary code developed by the histogram bins.

Multi-Resolution approach to grey-scale and rotation invariant texture classification based on LBP

This method is based on recognizing that certain LBP are fundamental properties of local image texture. This approach is very robust in terms of grey-scale variations since the operator is invariant against any monotonic transformation of the grey scale.

LBP based on Fourier transform

This method is a recent rotation invariant image descriptor computed from the discrete Fourier transforms of local binary histograms. The main idea of the method is based on constructing globally the invariants for the whole region to be described. It will retain the highly discriminative nature of LBP histograms.

There are several features that we could extract such as:

Local Binary Pattern Mean

$$LBPM = \sum_{j=1}^{h_{lbp}} p_j^{lbp} x_j^{lbp} \quad (3.21)$$

Local Binary Pattern Standard Deviation

$$LAS = \sum_{j=1}^{h_{lbp}} p_j^{lbp} (x_j^{lbp} - LBPM)^2 \quad (3.22)$$

Local Binary Pattern Entropy

$$LBPE = \sum_{j=1}^{h_{lbp}} -p_j^{lbp} \log p_j^{lbp} \quad (3.23)$$

3.1.3 Grey Level Co-occurrence Matrix Features

Grey Level Co-occurrence Matrix (GLCM) estimates image properties related to second order statistics. The GLCM is used to characterise texture patterns. The definition of GLCM can be described as a tabulation of how often different combinations of pixel brightness values i (grey levels) occur in a specific spatial relationship to a pixel with the value j . The number of grey levels in the image determines the size of the GLCM [10].

Let $P(i, j, d, \alpha)$ be the number of occurrence of pairwise grey levels i, j separated by a distance d and at a direction α , then the relative frequency corresponding to $P(i, j, d, \alpha)$ is

$$p(i, j, d, \alpha) = \frac{P(i, j, d, \alpha)}{N_{d, \alpha}} \quad (3.24)$$

where $N_{d, \alpha} = \sum_i \sum_j P(i, j, d, \alpha)$. An isotropic GLCM with $d = 1$ obtained from four matrices at $\alpha = 0^\circ$, $\alpha = 45^\circ$, $\alpha = 90^\circ$, and $\alpha = 135^\circ$, is

$$p(i, j) = \frac{1}{4} (p(i, j, 1, 0^\circ) + p(i, j, 1, 45^\circ) + p(i, j, 1, 90^\circ) + p(i, j, 1, 135^\circ)) \quad (3.25)$$

We can then obtain the estimated marginal probability from $p(i, j)$ as $p_x(i) = \sum_{j=0}^{N-1} p(i, j)$ and $p_y(j) = \sum_{i=0}^{N-1} p(i, j)$ where N is the number of distinct grey levels. Their means and standard deviations as μ_x , μ_y , σ_x , σ_y , respectively. The grey level difference histogram (GLDH) is defined as $D(k) = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} p(i, j)$, where $k = 0, 1, \dots, N-1$ and the grey level sum histogram (GLSH) as

$$S(k) = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} p(i, j), \text{ where } k = 0, 1, \dots, 2(N-1).$$

There are several features that can be extracted such as:

Contrast

Contrast measures the local variations in the grey level co-occurrence matrix and it is calculated using the following relation:

$$Cont = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} |i - j|^2 p(i, j) \quad (3.26)$$

Energy

Energy provides the sum of squared elements in the co-occurrence matrix. Energy expresses the angular second moment and it is calculated using the following relation:

$$Energy = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} p(i, j)^2 \quad (3.27)$$

Entropy

Entropy measures the uncertainty associated with a random variable. It is calculated using the information of the average matrix of the four grey level co-occurrence matrices. It is calculated using the following relation:

$$Entropy = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} -p(i, j) \log p(i, j) \quad (3.28)$$

Max Probability

Max Probability is the maximum element in the co-occurrence matrix and it is calculated using the following relation:

$$MaxProb = \max p(i, j) \quad (3.29)$$

Correlation

Correlation measures the joint probability occurrence of the specified pixel pairs and is calculated using the following relation:

$$\rho = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \frac{p(i, j)(i - \mu_x)(j - \mu_y)}{\sigma_x \sigma_y} \quad (3.30)$$

Diagonal Correlation

Diagonal Correlation measures the joint probability occurrence of pixels pairs in a diagonal form.

It is calculated using the following relation:

$$DiagCorr = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} p(i, j) |i - j| (i + j - \mu_x - \mu_y) \quad (3.31)$$

H_{xy1}

$$H_{xy1} = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} -p(i, j) \log(p_x(i)p_y(j)) \quad (3.32)$$

H_{xy2}

$$H_{xy2} = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} -p_x(i)p_y(j) \log(p_x(i)p_y(j)) \quad (3.33)$$

Difference Energy

Difference Energy provides the sum of squared elements of the grey level difference histogram and it is calculated using the following relation:

$$DEnergy = \sum_{k=0}^{N-1} D(k)^2 \quad (3.34)$$

Difference Entropy

Difference Entropy is calculated using the information of the grey level difference histogram and uses the following relation:

$$DEntropy = \sum_{k=0}^{N-1} -D(k) \log D(k) \quad (3.35)$$

Inertia

Inertia is a measure for the distribution of grey scales in the image. It is calculated using the information of the grey level sum histogram and uses the following relation:

$$Inertia = \sum_{k=0}^{N-1} k^2 D(k) \quad (3.36)$$

Homogeneity

Homogeneity measures the closeness of the distribution of elements in the co-occurrence matrix

to its co-occurrence matrix diagonal. It is calculated using the following relation:

$$Homogeneity = \sum_{k=0}^{N-1} \frac{D(k)}{1+k^2} \quad (3.37)$$

Sum Energy

Sum Energy provides the sum of squared elements of the grey level sum histogram and it is calculated using the following relation:

$$SEnergy = \sum_{k=0}^{2(N-1)} S(k)^2 \quad (3.38)$$

Sum Entropy

Sum Entropy measures the uncertainty associated with a random variable. It is calculated using the information of the grey level sum histogram and uses the following relation:

$$SEntropy = \sum_{k=0}^{2(N-1)} -S(k) \log S(k) \quad (3.39)$$

Sum Variance

Sum Variance measures the average of the squared differences from the Mean and it is calculated using the information of the grey level sum histogram. It is calculated using the following relation:

$$SVar = \sum_{k=0}^{2(N-1)} (k - \mu)^2 S(k), \text{ where } \mu = \sum_{k=0}^{2(N-1)} kS(k) \quad (3.40)$$

Sum Shade

Sum Shade measures the skewness of the matrix or in other words the lack of symmetry. It is calculated using the information of the grey level sum histogram and uses the following relation:

$$SShade = \sum_{k=0}^{2(N-1)} \frac{(k - \mu_x - \mu_y)^3 S(k)}{(\sigma_x^2 + \sigma_y^2 + \rho \sigma_x \sigma_y)^{3/2}} \quad (3.41)$$

Sum Prominence

Sum Prominence measures, as well as Sum Shade, the skewness of the matrix or in other words the lack of symmetry. It is calculated using the information of the grey level sum histogram and uses the following relation:

$$SProm = \sum_{k=0}^{2(N-1)} (k - \mu)^2 \frac{(k - \mu_x - \mu_y)^4 S(k)}{(\sigma_x^2 + \sigma_y^2 + \rho \sigma_x \sigma_y)^2} \quad (3.42)$$

3.1.4 Multiresolution Features

This set of features, as it is presented, is obtained from a nonseparable wavelet transform, the Quincunx Wavelet transform [26] [29]. A 2D quincunx wavelet transform is implemented with low and high pass filter banks similar to a 2D separable wavelet transform, the difference is that low and high-pass kernels can not be separated into two one-dimensional kernels.

In Quincunx downsampling, each subsequent low-pass image has half as many samples as its parent, a factor of $\frac{1}{\sqrt{2}}$ in each dimension.

Although separable wavelet transforms have a simple and well understood implementation, there are some considerations in using a non-separable wavelet decomposition: separable wavelet decompositions have vertical and horizontal cut-offs while the non-separable decomposition can have a cut-off of any angle, moreover, non-separable filter banks can be flexibly tailored for particular purposes, such having linear phases [30].

For features extraction only the first four even-level low pass decomposition images were retained.

Being $M = \frac{512}{k}$ the size (height or width) of the decomposition of the image at an even level L where $k = 2^L$ and $L = 1, 2, \dots, N$. N is the number of decompositions.

Mean

Mean is defined as the global average pixel value of the decomposition image at an even level L and it is calculated using the following relation:

$$M_k = \frac{1}{M^2} \sum_{i=1}^M \sum_{j=1}^M X_L(i, j), \quad (3.43)$$

Variance

Variance measures the pixel value variation and it is calculated using the following relation:

$$V_k = \sqrt{\frac{1}{M^2 - 1} \sum_{i=1}^M \sum_{j=1}^M (X_L(i, j) - M_k)^2}, \quad (3.44)$$

Skewness

Skewness measures the asymmetry on the tails of the distribution and it is calculated using the following relation:

$$S_k = \frac{1}{M^2 - 1} \frac{\sum_{i=1}^M \sum_{j=1}^M (X_L(i, j) - M_k)^3}{V_k^3}, \quad (3.45)$$

Kurtosis

Kurtosis measures the flatness or sharpness of the distribution and it is calculated using the fol-

lowing relation:

$$K_k = \frac{1}{M^2 - 1} \frac{\sum_{i=1}^M \sum_{j=1}^M (X_L(i, j) - M_k)^4}{V_k^4}, \quad (3.46)$$

Entropy

Entropy measures the uncertainty associated with a random variable and it is calculated using the following relation:

$$E_k = - \sum_{i=1}^{12} p_i^L \log p_i^L. \quad (3.47)$$

3.1.5 Gabor Features

Gabor features provide simultaneous localization in both the spacial and frequency domains. Using the frequencies of the Gabor filter-bank the features can be extracted [26].

Let an impulse response of Gabor filter be

$$g(x, y) = \frac{\exp \left[-\frac{1}{2} \left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} \right) + 2\pi j W x \right]}{2\pi \sigma_x \sigma_y}, \quad (3.48)$$

where W is the modulation frequency, x, y , are coordinates in the spacial domain, and σ_x and σ_y are the standard deviations in the x and y direction. A Gabor filter-bank consists of Gabor filters with Gaussians of different sizes modulated by a sinusoidal plane wave of different orientation from the same mother Gabor filter as

$$g_{m,n}(x, y) = a^{-m} g(\tilde{x}, \tilde{y}), a > 1, \quad (3.49)$$

where $\tilde{x} = a^{-m}(x \cos \theta + y \sin \theta)$, $\tilde{y} = a^{-m}(-x \sin \theta + y \cos \theta)$, $\theta = \frac{n\pi}{K}$, K is the number of the orientation, $n = 0, 1, \dots, K-1$, and $a = \left(\frac{U_h}{U_l} \right)^{\frac{1}{S-1}}$, where S is the number of scales, $m = 0, 1, \dots, S-1$ and U_h, U_l are the upper and lower center frequencies. The discrete Gabor filtered output of an image $I_E(r, c)$ of size $H \times W$ is given by a 2D convolution as

$$I_{g_{m,n}}(r, c) = \sum_s \sum_t I_E(r-s, c-t) g_{m,n}^*(s, t), \quad (3.50)$$

where $*$ indicates the complex conjugate. We obtain the mean and standard deviation of the energy of the filtered image,

$$\mu_{mn} = \frac{\sum_r \sum_c |I_{g_{m,n}}(r, c)|}{H \times W}, \quad (3.51)$$

$$\sigma_{mn} = \frac{\sqrt{\sum_r \sum_c (|I_{g,m,n}(r,c)| - \mu_{mn})^2}}{H \times W}. \quad (3.52)$$

Choosing the number of orientations and the number of scales for our Gabor filter-bank we can have a large number of features in the feature vector.

3.2 Classification Methods

Classification is a process of differentiating two or more classes by labelling each similar set of data with one class. This is a fundamental task in CAD systems and there are several methods of classification such as classification based on threshold, Support Vector Machines (SVM), Artificial Neural Networks (ANN), Decision Trees and K-nearest neighbour (KNN) method. There are two phases in the construction of a classifier. The training phase, where the training set is used to decide how the features should be weighted and combined, in order to separate the different classes, and the testing phase, where the weights determined in the training set are applied to a set of data that does not have known classes, in order to determine the class that this set of data should belong to [31].

3.2.1 K-nearest Neighbour

The KNN classification method is a non-parametric method that does not estimate parameters and that treats all the samples as points in the m-dimensional space, being m the number of variables. Given an input x, the proximity of neighboring observations in the training data set and their corresponding output values y are used to predict the output values in the validation data set and classify it according to the K-nearest training instances as determined by some distance metric, typically Euclidean distance. The KNN is known to be one of the simplest classifiers and requires these features to be distinctive enough for classification [32].

In 2012, Deepak *et al* [33] used the KNN classification method to classify a given ROI in benign and malignant after a learning set composed with only normal ROIs. A set of features were chosen from LBP and Radon Transform features. This approach should enable detection of the presence of lesions. If in a single mammogram several detected lesions exist, the classifier must classify each one of the lesions as benign or malignant before deciding if the given mammogram is of a normal or diseased breast.

3.2.2 Decision Trees

There are two types of decision trees: axis-parallel decision trees and oblique decision trees. In axis-parallel decision trees, a binary tree is built with each node having a single parameter that is compared to a constant. If the value of the feature is greater than the threshold, the right branch

of the tree is taken but, if the value of the feature is smaller, the left branch is followed. This procedure is repeated until one reaches a leaf node where all the objects are labelled as belonging to a specific class. These are called axis-parallel trees because they correspond to partitioning the parameter space with a set of hyperplanes that are parallel to all of the feature axes except for the one being tested. On the other hand, oblique decision trees allow the hyperplane at each node of the tree to have any orientation in parameter space, which means that at each node a linear combination of some or all of the parameters is computed and the sum is compared with a constant. The subsequent branching until a leaf node is reached is similar to axis-parallel trees. Oblique decision trees are more difficult in construction than axis-parallel trees because there are many more possible planes to consider at each tree node. This slows the training process down. Even though oblique decision trees are slower than axis-parallel decision trees, they are still faster to construct than neural networks. A major advantage of decision trees is the interpretability; additionally, they often produce very simple structures relying on only a few parameters to classify the objects [31].

Liu *et al* [34] in 1999, introduced in their work, the Binary Tree Classifier (BTC) a method to classify features at multiple resolutions: fine and coarse. Liu *et al* used an iterative growing and pruning algorithm for the classification tree design because it produces trees with higher classification accuracy and requires less computation. The authors used only a sub-sampling of the training set to grow the BTC and after it was generated, they associated with each terminal node a probability of suspicion, which is the percentage of lesions pixels in the training images that falls in this terminal node. The suspicious probability is then recomputed using the entire set of training samples.

Sun *et al* [35] also used BCT to classify normal mammograms. In 2002, the authors used a BTC to classify ROIs as normal regions or suspicious regions, independently of abnormalities present in the region. Twenty features were extracted from Multi-resolution features and used to train and test the Binary Decision Tree that was constructed using the convergent algorithm. Some fairly satisfactory results were obtained although some cancers failed to be identified. After using BCT classifiers, Sun *et al* [36] also used a cascading classifier, which means that they used a combination of BCT classifier with a linear classifier in their works. Using a set of 86 features extracted from curvilinear features, GLCM features, Gabor features and multi-resolution features, they proposed a multi-stage cascading classifier in order to improve the classification performance. This method was proposed due to the fact that a single classifier may not be sufficient to differentiate each sub-pattern. The multi-stage cascading classifier was expanded from the two-stage classifier, where, after the classification through a BCT in normal and abnormal class, the abnormal class is classified, one more time, with a linear classifier using Fisher's criterion and 11 of the initial 86 features. The first stage correctly classify most of the abnormal regions while separating out as many of the normal regions as possible. In the second stage, the BCT is be used to classify as normal or abnormal, those images that were classified as abnormal in the first stage. It was proved by the authors that this method improves not only the true negative fraction, but also the correct classification rate of normal mammograms while keeping a low false positive fraction and

improves also the misclassification rate of abnormal mammograms as normal [37].

A particular case of decision tree classifiers, that is also used in articles reviewed in this state-of-the-art, is the method that uses threshold as classifier.

In 1994, Taylor *et al* [38] proposed a method capable of separate "easy" and "difficult" mammograms. This method was one of the first methods towards automated systems for identification of normal or abnormal mammograms. The proposal was to classify mammograms into dense or fatty using three different approaches: two experienced radiologists (which offered a reliable comparison method for the automated system), the measures of the image texture and finally the automated procedure. In the second method, a set of local statistical and texture features were tested in terms of separation between dense and fatty mammograms, where it was concluded that local skewness was a good feature to separate mammograms. In the automated procedure, the method consisted in using a simple global threshold to segment the background from the foreground and incorporate a set of features that give a good separation between dense and fatty tissues. These features have been tested in the previous method. This method separates approximately two thirds of the fatty mammograms.

3.2.3 Random Forests

Random Forests (RF) are a bootstrapping method applied to classification trees. It grows several classification trees, thus the name Forests. Each classification tree is a complete tree with leaves that consist of some pre-selected number of observations. When classifying a new object, the input vector associated with this new object is processed by each one of the trees. Each tree gives a classification and in the end the forest chooses the classification having the majority of votes. The classification trees could be grown in 3 different forms:

1. Supposing the number of cases in the training set is N , N random sample cases are selected (with replacement) from the original data. This sample will be the training set for growing the tree.
2. Having M available variables in the input vector, m variables between the M available are selected randomly. The best split on these m is used to split the node. During the forest growing process, the number of variables m is held constant.
3. The trees grow to the largest length possible without associated pruning.

After the construction of each tree, all of the data are run down the tree, and for each pair of cases the proximities are computed. If for any reason, two different cases occupy the same terminal node, their proximity is increased by one unity. In the end, the proximities are normalised.

The error rate of the RF depends on the correlation between any two trees of the forest and on the strength of each individual tree. Increasing the correlation between trees increases the forest error rate in contrast with the strength, that when increased provides a decreasing of the forest error rate. If the number of variables m is reduced, both correlation and strength are also reduced. On the

other hand, if the number m is increased, both correlation and strength are increased. Somewhere in the middle, there is an optimal range of m . To determine the optimal range of m , the Out-of-bag (Oob) rate is used. Oob rate is the only adjustable parameter to which RF are sensitive to [39].

3.2.4 Artificial Neural Networks

A neural network consists of units, arranged in layers, which convert an input vector into some output. Each unit takes an input, applies a function to it and then passes the output on to the next layer. Feed-forward Neural Networks (FFNN) uses a unit that feeds its output to all the units on the next layer, but there is no feedback to the previous layer. In the learning phase, weights are applied to the signals passing from one unit to another, and it is these weights that are tuned in the training phase to adapt a neural network to the particular problem that one is facing. The main advantage of this methods is that they generalize well, which means that they can handle problems with many parameters, and they are able to well classify objects even when the distribution in the N -dimensional parameter space is very complex. The disadvantage of neural networks is that they are considerable slow [31].

In 1996, Sahiner *et al* [40] used a Convolution Neural Network (CNN) to classify Regions of Interest (ROI) on mammograms as either normal tissue or masses. He used a three-layer CNN with three input images derived from ROI's using two different approaches: i) averaging and sub-sampling and ii) Spatial Grey Level Dependence (SGLD) and Grey Level Difference Statistics (GLDS) features extracted and applied to small regions inside the ROI, by themselves or in a combination of both. The results of the work indicated are that the choice of CNN input images is more important than the choice of CNN architecture, as well as the combination of the outputs from CNN classifier and multi-resolution texture analysis classifier improves classification accuracy.

Kalman *et al* [41] used, in 1997, a combination of wavelet transform and Artificial Neural Network (ANN) to detect mammograms with a mass. The vectors composed by the wavelet were used to train the FFNN to classify the image pairs for the presence of a mass. The FFNN was composed by six input units, four hidden units and five output units divided in two groups: categorization units and hint units containing the indication of location of a mass. The authors concluded that this combination is promising and may provide a reliable method to pre-screen mammograms for masses with high sensitivity and reasonable specificity.

A most recent work using ANN was the work of Shinde in 2003 [15], where she aimed to segment breast masses from the normal background breast tissue. 198 different features were extracted using manual segmentation. Shinde used a quick propagation neural network with the leave-one-out validation method to classify the masses as benign and malignant, based on a set of extracted features. Shinde considered that the success of automated classification requires the knowledge of the mass, the ambient normal-tissue, the background border region and the tumor area.

3.2.5 Support Vector Machines

SVM are based on the concept of decision planes that define decision boundaries. A decision plane separates between a set of objects having different class memberships. There are two main types of SVM classification methods: binary SVM classifiers and, because most of the problems of the real life are not binary, multi-class SVM classifiers were developed. Multi-class SVM classifiers can be of several types such as one-versus-rest, one-versus-one, DAGSVM, method of Weston and Watkins [42] and the method of Crammer and Singer [43]. A more detailed explanation of the different multi-class methods is presented in Statnikov *et al* work [32]. SVM supports both regression and classification tasks and is efficient enough to handle very large-scale classification in both sample and variables. They often achieve superior classification performance compared to other algorithms.

In 2004, Chiracharit *et al* [44], used SVM classifiers to classify mammograms in normal or abnormal ones. A group of 90 crossed distribution pairs involving 54 features are identified and mapped into feature space. The SVM was constructed with crossed feature pairs, used to train the SVM classifier. The authors also used separate SVMs classifiers with different kernel functions for each feature class and the final normal or abnormal decision was made by majority vote. With this work, the authors have been able to conclude that the method extracts useful information using the SVM approach that combined with information from ordinary features, improves the automated classification of mammograms. Following the previous work developed by the authors, in 2007 they proposed a Local Probability Difference-uncrossed-feature based SVM learning system [26] that is the combination of uncrossing mapping and local probability difference transformed features that could fix the "crossed distribution"¹ and "noise" problems. This method can be used with normal mammogram features more effectively and tries to improve the specificity from the previous methods, while keeping a high sensitivity at the same rate.

Elshinawy *et al* was another group of authors that used SVM classifiers in their works of 2010 and 2011. They used one-class SVMs classifiers and two-class SVMs classifiers. One could say that the authors work is distinguished from work of other authors by the previous separation of the mammograms into dense or fatty tissue. This method was firstly mentioned by Wolfe [45] that introduced the idea of breast density classification. The authors used in their works different sets of features and both one and two-class SVM classifiers or simply one-class SVM classifier towards the evaluation of the performance of the method proposed. In their early work they used only one-class SVM classifiers to evaluate the effect of the density in selecting features for normal mammogram detection, where they used GLCM features and LBP features and a method towards normal detection using only three different sets of LBP features and one-class SVM classifier. They were able to conclude with this approach that, GLCM features have a better performance when extracted from mammograms with fatty tissue while LBP features are more efficient when extracted from mammograms with dense tissue [46] [27]. In three of the authors works, they used

¹problem originated by the confounding distributions of the features for both normal and abnormal cases

one-class and two-class SVM classifiers, in order to compare them towards normal mammogram detection. The one-class SVM classifier was trained with only normal mammograms whereas the two-class was trained with both normal and abnormal mammograms. In one of the works, they used a set of 13 GLCM features where they have showed that first and second-order statistics based on GLCM features, when used together, outperform using only first order statistics. They compared also the use of both GLCM and LBP features for one and two-class SVM classifiers in Elshinawy PhD thesis. The results of the authors work shows that one-class SVM outperforms two-class SVM for normal detection and that the false negative rate was significantly reduced when using both LBP and GLCM [47] [10] [48]. In all of the previous mentioned studies, the performance of the classifier was evaluated on the performance of each block and to evaluate the final performance of a full-field mammogram it was used majority voting approach. Finally, it was possible to conclude that this new approach, taking into account the density of the tissue in the mammogram improves the classifier performance.

Most recent work using SVM classifiers was presented by Domingues *et al* [49] in 2012, where the authors used SVM classifiers to classify the mammograms in two phases. First, after extracting 9 features from only the CC view of the mammogram, it was classified in terms of density in dense or non-dense mammograms. After this first phase, a new set of features composed by 5 GLCM features and a last feature from LBP image were extracted from only the non-dense mammograms that were classified one more time but, in this case in terms of malignancy, being benign or malignant mammograms. Both SVM classifiers used were created with an RBF kernel function.

3.2.6 No Classification Method

Some articles do not use a classification method in their proposed methodology for identification and characterization of the normal tissue.

Liu et al [50] in 1998 proposed a normal tissue identification and removal algorithm, independently of the types of abnormalities that might be present in the mammogram. The proposed method consists in smoothing the image with a Gaussian lowpass filter and then, subtract the smoothed image from the original image to achieve a image with simply the background. This new background image contains normal linear markings enhanced while the background densities and abnormal masses were largely removed. The method proceeds with a "normal" structure detection and analysis that continues with the elaboration of a mask where all the pixels in the original mammogram are set to zero. This proposed method was the first step in the elaboration of an early CAD system.

Addaoudi *et al* [51] in 2005, proposed a method capable of characterizing the various tissues constituting a normal mammogram. For this purpose, the authors extracted a set of GLCM features such as average, variance, energy, contrast and correlation. After this extraction, unlike other authors works, there is no classification method, but the features are applied to both fat and fibrous tissue. A further investigation of the results of these works, allows to conclude that features such as variance, contrast and correlation allow to distinguish between the fat and fibrous tissue.

3.3 Density Classification

Several works have mentioned the effect of density when talking about BC, CAD and Pre-CAD systems.

Wolfe [52] [45] was the first author mentioning the idea of breast density classification in 1976. The author used solely visual classification of mammograms to classify them according to the density into four categories: N1 (Normal Breast), P1 (Mainly Fat), P2 (Involvement of the breast with a prominent duct pattern) and DY (Mainly Dense). He concluded that a relation between the four categories used to classify the breast and the risk of developing BC exists. As the density in the breast increases, also the risk of developing BC increases, thereby it was possible to conclude that N1 is the category with Lowest Risk of developing BC and DY the category with Highest Risk of developing BC.

The idea that there is a relation between the breast density and the risk of developing BC was followed in 1995 by *Boyd et al* [53], where the authors used a quantitative computerised thresholding method in order to measure the density of the breast. This method is based on the proportion of dense breast tissue relative to the breast and has six categories according to the percentage of density: Class 1 for 0% of density in the breast, Class 2 for a percentage of density between 0 and 10%, Class 3 for a percentage of density between 10 and 25%, Class 4 for a percentage of density between 25 and 50%, Class 5 for a percentage of density between 50 and 75% and Class 6 for a percentage of density between 75 and 100%.

In 2002, *Bovis and Singh* [54] investigated a new approach in the classification of mammograms according to the tissue type using a combined classifier paradigm. They have studied this task by following two different paths: the four-class classification problem and the two-class problem. The two-class problem classifies the tissue in one of two classes: fatty or dense. In other hand, the four-class approach classifies the tissue in fatty, partly fatty, dense and extremely dense. The results obtained by *Bovis and Singh* indicated that the classification according to the BI-RADS is a complex task and increases the chance of confusion in the resultant classification. In opposition to this, the two-class classification using this new method proposed by *Bovis and Singh* improves the performance and robustness of the classification. Finally, the authors concluded that the results obtained justify the use of this method in a CAD system.

Following the study of density and CAD systems and, as the results obtained previously in the classification according to the BI-RADS were not satisfying, *Petroudi et al* [55] investigated how the specificity and sensitivity of CAD systems can be affected when these systems are used only for the assessment of mammograms that belong a specific BI-RADS category. The authors concluded that the system performed better if applied to breasts belonging either to the first or second BI-RADS classes. Taking into account the density of the breast in CAD systems can significantly improve their performance in terms of sensitivity and specificity for at least a certain number of classes.

In 2008 *Oliver et al* [56] proposed a novel automatic breast tissue density classification method. In a first phase, the breast area is segmented using the grey-level information. Subse-

quently, the fuzzy C-means algorithm is used to segment fatty versus dense tissue types. For each tissue region, features are extracted to proceed to the characterization of the breast tissue and finally the mammograms are classified, using a combination of classifiers, according to BI-RADS categories.

All the above mentioned works have the density as basis of research, since the early study of breast density versus risk of BC till the modern CAD systems that act as a "second-look" in the mammogram detection. Another important area in this field is the area of pre-CADs that act as a "first-look" in the detection of BC. In the area of pre-CADs, many authors have done research towards normal detection, but so far only a few have introduced the density classification as the first step in normal mammogram detection. Authors such as *Elshinawy et al* [10] and *Domingues et al* [49] have done some research in the specific field of normal mammogram detection based on tissue type. *Domingues et al*, unlike *Elshinawy et al*, sends the dense mammograms directly to expert evaluation and classifies only the fatty ones, fact that causes the abnormal mammograms to be lost in the process.

In 2010 *Elshinawy et al* [27] proposed an algorithm to separate the mammograms into dense and fatty categories and to extract features from each category individually and finally proceed to a normal mammogram classification using one-class SVM classifier. The authors concluded that it is possible to obtain a better accuracy using majority voting with rotation invariant LBP features, rather than using basic LBP operator features. In the following work of the authors [46], rather than using only LBP features, they used also GLCM features with only first-order statistics and GLCM features with both first-order and second-order statistics. The authors concluded that both GLCM with only first-order statistics and GLCM with both first and second-order statistics were more accurate than basic LBP and rotation invariant LBP features. Using GLCM features with both first and second-order statistics improves so much the system, that they managed to reach a rate of FN of 0% and a rate of TP of 100% for both dense and fatty mammograms. For dense mammograms they reach a rate of TN of 96.31% and a rate of FP of 3.69%. For fatty mammograms the rates decrease a little in terms of TN to 95.85% and increases in terms of FP to 4.14%. Both the previous referred works of *Elshinawy* have presented results from her PhD Thesis. In this work [10], the author not only studied the use of LBP and GLCM features in the mentioned different forms but also extracted Gabor features in order to compare the classifier performance where the author obtains values for the recall rate between 97.34% and 99.75%. With all this works, the author was able to conclude that separating the dense and fatty mammograms reduces the FN rate in each tissue type individually when GLCM and LBP features were extracted. For the FP rate, LBP based on Fourier Transform performed better with dense mammograms while GLCM features performed better for fatty mammograms. Gabor features did not show promising results.

In 2012 *Domingues et al* [49] used a two block method to classify normal mammograms. In first phase the authors classified the mammograms into dense and fatty and extracted GLCM and LBP features from only fatty mammograms since dense mammograms were immediately sent to expert evaluation. In a second phase they used SVM classifiers created with an RBF kernel

function to classify the mammograms one more time but, in this case in terms of malignancy, being benign or malignant mammograms. The authors used a total of 410 images from the INbreast database [57] composed by 283 fatty images and 127 dense images. The database was divided into train and test sets (75%|25%). The density block classifier presents an accuracy of 50% and FN rate of 6%. The benign classifier block was trained with 26 benign images and 23 malign images previously classified as fatty and achieved a TP rate of 12%, a TN rate of 28%, a FP rate of 61% and a FN rate of 0%.

3.4 Summary

Although plenty of research has been done in this field, as it was possible to see in this state-of-the-art chapter, there is still no fully satisfactory solution.

The work to be proposed follows the main ideas of *Elshinawy et al* [10] and *Domingues et al* [49], even if the work of *Elshinawy et al* [10] fits better in what we want to develop. This happens because the author extracts features from both dense and fatty mammograms and performs normal detection without ignoring abnormal mammograms. This is the aim of the proposed work and will be explained in detail in the next chapter.

Chapter 4

Methodology

This chapter presents the methodology adopted in the realization of this MSc Dissertation. The methodology was based in four major steps: Density Classification, Pre-Processing of the mammograms, Features Extraction and Normal/Abnormal Classification. The block diagram with the approach that was used is presented in Figure 4.1.

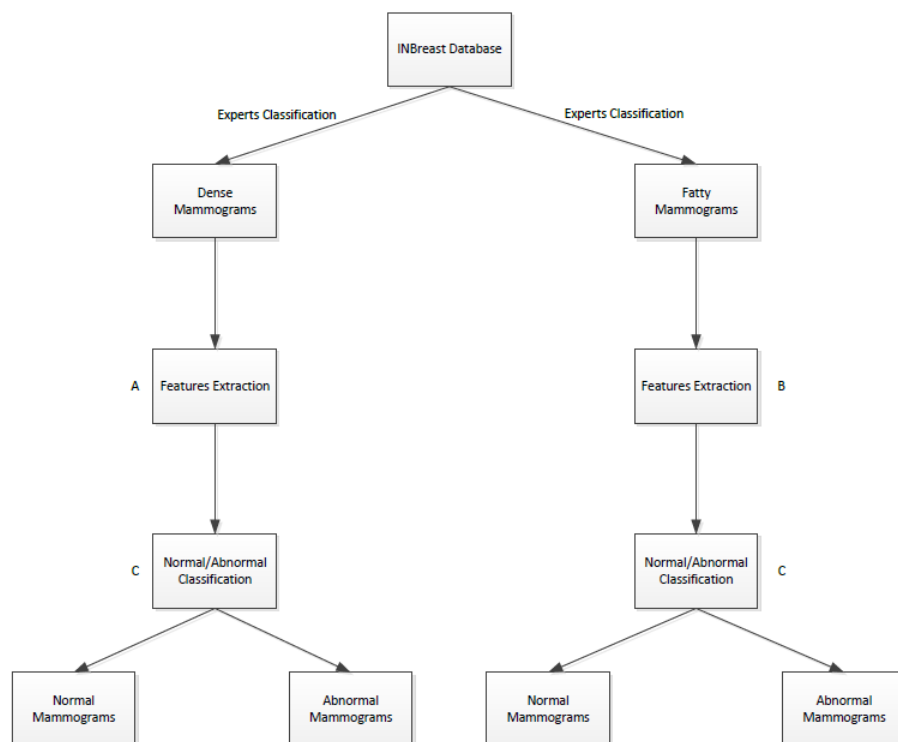


Figure 4.1: General Block Diagram of the Proposed Methodology: A, B - GLCM; C - SVM, Random Forests, K-nearest neighbour

Having our database, it will be divided into two mammogram sets, the dense mammograms and the fatty mammograms. From each one of the sets, will be extracted two types of features, the GLCM features and the LBP features. After having files with each one of the features for each

tissue type, they will be used in the classification phase. This classification phase will happen in in three different ways, only GLCM features, only LBP features and GLCM and LBP features together. This classification will tells us if the mammogram is normal or abnormal.

4.1 Database

In this work, it was used mammograms with only the CC view, from the INbreast database [57], a recently proposed database. The images of this database were acquired at the Breast Center in Centro Hospitalar São João with MammoNovation Siemens Full-field Digital Mammography (FFDM) equipment and saved in Digital Imaging and Communications in Medicine (DICOM) format¹. INbreast has FFDM images from screening, diagnostic and follow-up cases. A total of 115 cases were collected. 90 of the cases have two images (CC and MLO) of each breast and the remaining 25 cases are from women who had a mastectomy and two views of only one breast were included. The images were distributed according to the density for each BI-RADS class being BI-RADS 2 the class with more cases (more than 50% of the total number of cases). In almost all the BI-RADS classes, the number of fatty mammograms is higher than the number of dense mammograms [57]. The used mammograms have been previously classified by experts, into dense and fatty mammograms, in order to be possible to study the efficiency of the use of different sets of features combined with several classification methods.

The distribution of the mammograms according to the BI-RADS class can be seen in Figure 4.2.

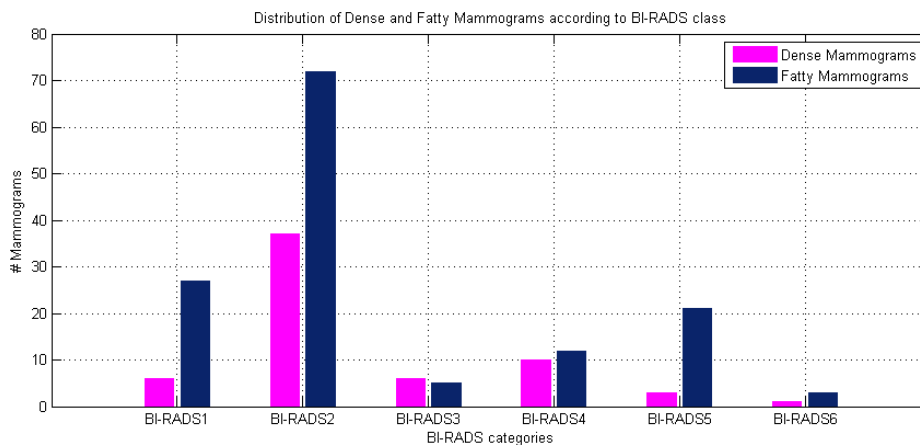


Figure 4.2: Distribution of Dense and Fatty Mammograms according to BI-RADS class

A total number of 203 mammograms was used. Those mammograms were divided into 63 dense mammograms and 140 fatty mammograms. In the first tested scenario BI-RADS categories 1, 2 and 3 were considered benignant and the BI-RADS categories 4, 5 and 6 were considered

¹DICOM format gathers the image and also some related metadata

malignant. In a second phase, it was opted to consider a new scenario where the BI-RADS categories 1 and 2 were benignant categories and BI-RADS categories 3, 4, 5 and 6 were malignant categories.

Of this total number of mammograms, in the first scenario, 153 mammograms were considered normal and 50 were considered abnormal mammograms, on the other hand, in the second scenario, 142 were considered normal mammograms and 61 mammograms were considered abnormal. The distribution of the mammograms in the dataset, according to the second scenario can be seen in Table 4.1. It was opted to consider this second scenario in the tests, because it is the scenario that privileges the detection of malignant mammograms, the main purpose of this work. The first scenario, although some preliminary tests have been done in the beginning of this work, in order to compare both scenarios, the first one has been abandoned in the realization of further tests.

Table 4.1: Distribution of Dataset according to the second scenario

	<i>Normal</i>	<i>Abnormal</i>	
<i>Dense</i>	43	20	63
<i>Fatty</i>	99	41	140
<i>TOTAL</i>	142	61	203

4.2 Pre-Processing of the Mammograms

Pre-processing was applied to all mammograms in order to remove possible irrelevant artefacts and reduce the dimensionality of the mammograms. As the background of the mammograms was well defined with a threshold equal to zero, it was not necessary to proceed to the segmentation of the breast in the mammogram, although this was a scenario that was taken into account when the possible problems that could arise in the development of this dissertation were studied.

Five major steps have been considered in the pre-processing phase:

1. **Contrast and brightness adjustment** in order to enhance the mammogram.

For the adjustment it was used the Matrix Laboratory (MatLab) function *imadjust* [58]. In this case, it was used the form of the function where it is possible to adjust the input parameters and the output parameters.

$$J = \text{imadjust}(I, [\text{low in}; \text{high in}], [\text{low out}; \text{high out}])$$

The function maps the values in the original image to new values in the new image created such that values between the input values map to values between the output values.

The values for *low in*, *high in*, *low out*, *high out* must be between 0 and 1. It was used, for this work, a *low in* and a *low out* equal to 0, an *high in* equal to 0,05 and an *high out* equal to 0,9.

2. Determination of mammogram's laterality

To determine if the mammogram is from a left or right breast, it was used a method that uses histogram analysis (by comparing the average intensity level of the right half with the left half) [59].

If the average intensity of the right half is higher than the average intensity of the left half, the mammogram is from a right breast. On the other hand, if the average intensity of the left half is higher than the average intensity of the right half, the mammogram is from a left breast.

3. Determination of the most extreme point

In mammograms containing the right breast, the most extreme point is the most left point, with threshold different than zero. On the other hand, in mammograms containing the left breast, the most extreme point is the most right point, with threshold different than zero.

To determine the most extreme point it was tested the threshold value, horizontally, for each pixel of the mammogram.

In right mammograms, the coordinates of the pixel that is being tested were saved only if the coordinates of that pixel, in terms of the x-axis, were lower than the last saved coordinates and the threshold of that pixel is higher than zero. In left mammograms, if the threshold of the pixel that is being tested was higher than zero, the coordinates of the pixel were saved if those coordinates, in terms of the x-axis, were higher than the last saved coordinates.

4. Determination of the highest and lowest point

The process to determine the highest and lowest point with threshold different of zero in the mammogram is similar to the process to find the most left/right point. The difference between them was that, instead of saving the coordinates of the pixels when the coordinates of the x-axis are higher/lower than the last saved coordinates, the coordinates of the pixels are saved when the coordinates of the y-axis are both higher or lower than the last saved coordinates.

5. Resizing of the mammogram according to the three points previously determined. In the resizing of the image it also takes into account that the final size should be such that it is possible to divide the image into 512×512 blocks.

In Figure 4.3 it is possible to see an original left breast mammogram, the same original mammogram with the most right, higher and lower points, with threshold different to zero, marked and the resized mammogram.

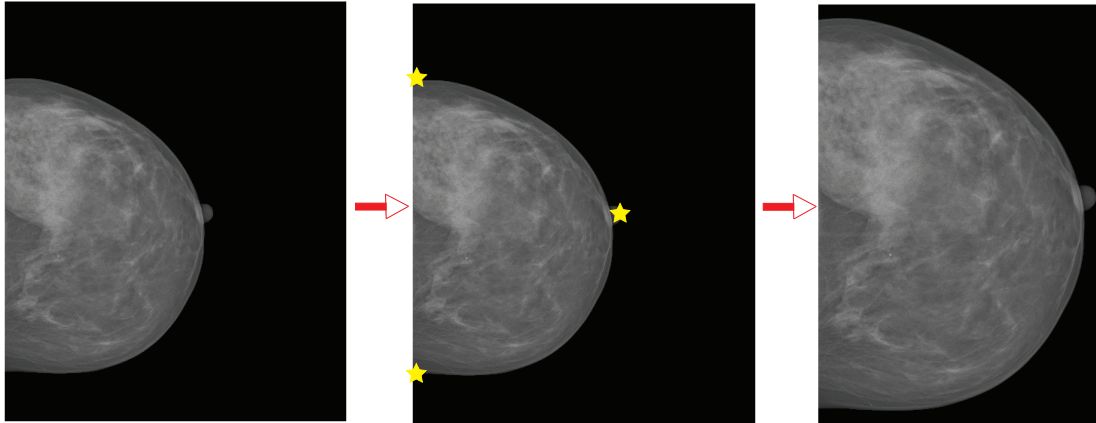


Figure 4.3: Determination of most high, low and extreme points

4.3 Features Extraction

In this work, GLCM and LBP features were used and three different case scenarios were considered. In the first scenario, as Elshinawy *et al* concluded in [46] [27], that a better performance is accomplished when different sets of features are extracted from different types of mammograms, GLCM and LBP features were extracted separately from both type of mammograms. In another scenario, both GLCM and LBP features were used together in the classification phase. Finally, in the last scenario, the density of the breast was not taken into account and GLCM and LBP features were extracted from all mammograms and used separately in the classification phase.

4.3.1 GLCM Features

As said in 3.1.3, GLCM can be described as a tabulation of how often different combinations of pixel brightness values i (grey levels) occur in a specific spatial relationship to a pixel with the value j . The number of grey levels in the image determines the size of the GLCM [10].

In this work, to obtain the GLCM matrices it was used a distance $d = 1$ and four different directions for α : $0^\circ, 45^\circ, 90^\circ, 135^\circ$.

Five features were extracted from each one of the GLCM matrices and for each one of this features, it was calculated the mean and standard deviation:

- **Contrast** measures the local variations in the grey level co-occurrence matrix.
- **Correlation** measures the joint probability occurrence of the specified pixel pairs.
- **Energy** provides the sum of squared elements in the co-occurrence matrix.
Energy expresses the angular second moment.
- **Homogeneity** measures the closeness of the distribution of elements in the co-occurrence matrix to its co-occurrence matrix diagonal.
- **Entropy** measures the uncertainty associated with a random variable. It is calculated using the information of the average matrix of the four grey level co-occurrence matrices.

Thus, a total number of ten features were extracted from the four GLCM matrices and saved in a file, to be used later in the classifier.

To determine the four GLCM matrices, we used the MatLab function *graycomatrix* that determines the GLCM matrices according to the Offset parameter specified in the function. This parameter also includes information about the distance between the pixel of interest and its neighbour, as shown in the formula for $0^\circ, 45^\circ, 90^\circ, 135^\circ$ respectively:

$$Offsets = [0 \ d; -d \ d; -d \ 0; -d \ -d] \quad (4.1)$$

GLCM represents the distance and angular spatial relationship over an image sub-region of specific size.

The *graycomatrix* function calculates the GLCM matrices having just a grey-scale image. The calculations are made so it is possible to know how often a pixel with a certain grey-value i occurs adjacently to pixels with a value j in each one of the previous mentioned directions $0^\circ, 45^\circ, 90^\circ, 135^\circ$, as represented in Figure 4.4.

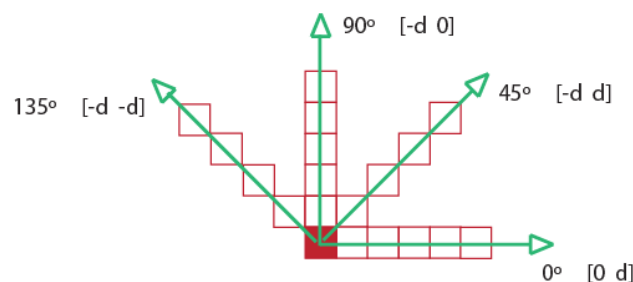


Figure 4.4: Scheme directions calculations GLCM

The i and j variables are the grey-values in the image. The dimensionality of the GLCM matrices depends on the number of grey-values that the image contains: 8 or 256. In this work, the GLCM matrices produced for each of the directions had 256×256 because the image had 256 levels of grey. In Figure 4.5 an example of how the GLCM matrices are produced is presented, but for only 8 levels of grey [60].

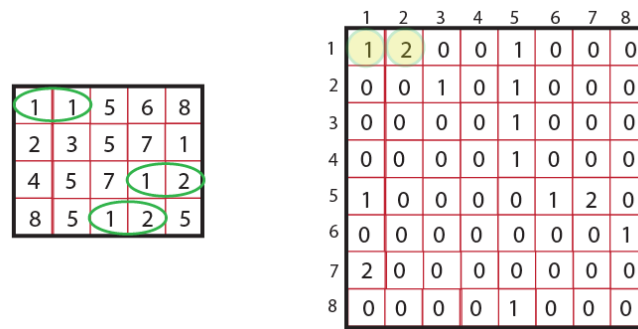


Figure 4.5: Scheme calculation GLCM (8 levels) for 0° or $[0 \text{ d}]$

The number of grey levels in the image is also specified on the parameters of the function, in this case the number of levels was 256, which means that the function scales the values in the image so they are integers between 1 and 256. The parameter symmetric was set to true, which means that the function counts both $1,2$ and $2,1$ pairings when calculating the number of times the value 1 is adjacent to the value 2.

Although it exists a MatLab function for extracting the features, *graycoprops*, the function extracts only four features of the five wanted features: contrast, correlation, energy and homogeneity. Due to this fact, a function for extracting the features was created based on *graycoprops* but with an extra feature, the entropy.

The first step when creating the function to extract the features from the GLCM matrices was to normalise the four matrices, so that the sum of its elements is equal to 1. The formulas of the features that we want to extract, defined in the 3.1.3, were applied and the features for each matrix were extracted. After the extraction of the five features for each matrix, the mean and standard deviation were calculated and normalised to have zero mean and unit variance, and the total number of ten features was saved in a file to be used later in the classification phase.

4.3.2 LBP Features

In this work two different set of features based on LBP were extracted. The basic LBP and a multi-resolution approach to grey-scale LBP. For the first set of features, the basic LBP, a function based on the study introduced by Ojala *et al* [61] was implemented.

For each 3×3 block of the image, the center was encountered and the threshold of the central pixel determined. Then, the threshold of the central pixel was compared with the rest of the neighbours of the block. If a neighbour had a higher or equal grey level than the pixel of the center, the value 1 was attributed. On the other hand, if the grey level of a neighbour was lower, the value 0 was attributed to the neighbour.

The scheme of the basic LBP, as it was explained, can be seen in Figure 4.6.

The LBP code was produced by concatenating the ones and zeros obtained previously into a binary code. Finally the binary code was transformed into a decimal code by multiplying the

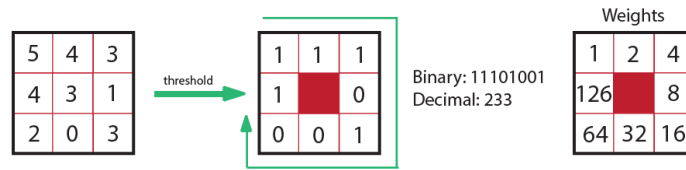


Figure 4.6: Scheme of basic LBP

binary code with the weights represented in the Figure 4.6. The decimal code is used to further analysis through an 256 bins histogram, as the block is 3×3 . Through the analysis of the histogram, five features were extracted:

- **Mean** is defined as the global average pixel value of the image.
- **Standard Deviation** measures the dispersion in a frequency distribution.
- **Entropy** measures the uncertainty associated with a random variable.
- **Energy** provides the sum of squared elements in the histogram.
Energy expresses the angular second moment.
- **Skewness** measures the asymmetry on the tails of the distribution.

The extracted features were saved in two files. One of the files containing only two of the features: mean and standard deviation, and the other file with the total number of features. This two files were saved with different sets of features to be possible to study which is the influence of some features, which means, study if having more features can benefit the classification of mammograms. These files were used posteriorly in the three different classifiers.

In the case of the multi-resolution approach to grey-scale, it was used a code implemented by Marko Heikkilä and Timo Ahonen [62] [63] for both $LBP_{8,1}$, 8 neighbours and radius 1, and $LBP_{16,2}$, 16 neighbours and an radius of 2. The Figure 4.7 illustrates the two different neighbourhoods used in this work.

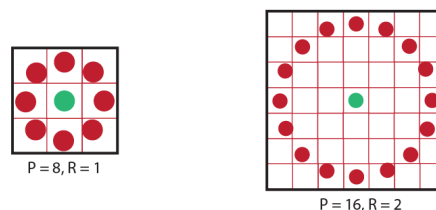


Figure 4.7: Scheme of Multi-Resolution approach to grey-scale LBP

This new approach to LBP features is an improved approach of the basic LBP. Instead of having a squared neighbourhood of 3×3 , we have neighbourhoods of different sizes. From the central pixel of the neighbourhood, a circle with radius R is made having P sample points to compare with the threshold of the central pixel, like in the Basic LBP case.

Being (x_c, y_c) the coordinates of the central pixel, to determine the coordinates of the sampling points P on the edge of the circle with radius R, the following formulas should be used:

$$x_p = x_c + R \cos(2\pi p/P) \quad (4.2)$$

$$y_p = y_c - R \sin(2\pi p/P) \quad (4.3)$$

Let g_c be the grey level of the central pixel and the grey levels of his neighbours are g_p , with $p = 0, 1, \dots, P-1$. The local texture of the image is characterized by the joint distribution of grey values of the points on the circle, $p = 0, 1, \dots, P-1$:

$$T = t(g_c, g_0, g_1, \dots, g_{P-1}) \quad (4.4)$$

The central pixel can be subtracted from the neighbour pixels without losing information about the local texture of the image:

$$T = t(g_c, g_0 - g_c, g_1 - g_c, \dots, g_{P-1} - g_c) \quad (4.5)$$

Assuming that the central pixel of the neighbourhood can be statistically independent of the differences, the texture T is factorized:

$$T \approx t(g_c) t(g_0 - g_c, g_1 - g_c, \dots, g_{P-1} - g_c) \quad (4.6)$$

Now the function is composed by two factors and both factors have different importances. The first factor $t(g_c)$ is the intensity distribution over the image and as it is not related to the local image texture, it does not provide useful information for texture analysis. On the other hand, the other factor, the joint distribution of differences can be used to model the local texture:

$$t(g_0 - g_c, g_1 - g_c, \dots, g_{P-1} - g_c) \quad (4.7)$$

To achieve invariance with respect to grey scale, only the signs of the differences $g_{P-1} - g_c$ are considered, which means that in case a point on the circle has the same or a higher grey value than the center pixel, a one is assigned to that point, else a zero is assigned to that point.

$$t(s(g_0 - g_c), s(g_1 - g_c), \dots, s(g_{P-1} - g_c)), \quad (4.8)$$

where

$$s(z) = \begin{cases} 1 & z \geq 0 \\ 0 & z < 0 \end{cases} \quad (4.9)$$

The LBP for the pixel (x_c, y_c) , is produced assigning to each $s(g_p - g_c)$ a weight 2^p , similarly to Basic LBP.

These weights are summed:

$$LBP_{P,R}(x_c, y_c) = \sum_{p=0}^{P-1} s(g_p - g_c) 2^p \quad (4.10)$$

Thus, the local binary pattern characterizes the texture around (x_c, y_c) .

Having this operator with a neighbourhood with $P = 8$ and $R = 1$ is very similar to having a Basic LBP operator with a neighbourhood of 3×3 [63].

4.4 Classification

Classification is one of the most important tasks when developing a pre-CAD or CAD system, as well as in other learning tasks.

As mentioned in 3.2, classification is a process of differentiation between two or more classes by labelling each similar set of data with a specific, corresponding class. It is known that a classifier has two different phases in its construction: the training phase and the testing phase.

In terms of classification methods, it was used, for each one of the three different scenarios explained in 4.3, three different classification methods: K-nearest neighbour, Support Vector Machine and Random Forests. The employment of these three different scenarios was made to understand which of these three classifiers is more efficient in normal/abnormal classification of mammograms when combined with a determinate set of features.

To compare the classifiers and estimate the performance we used the leave-one-out method. This method is more suitable for the type of sets that we used in this work, small datasets.

4.4.1 K-Nearest Neighbour

The K-Nearest Neighbour classifier was chosen to take part in this work because it is a very simple algorithm for both understand and implement. Thus, this is an algorithm that should always be considered in classification tasks, although is also an algorithm that requires large memory and it is a bit slow, due to the fact that it is necessary to calculate, for each testing instance, the distance to each one of the training instances, making the prediction slower than desired.

Another of the disadvantages of the KNN is that the accuracy of the algorithm can degrade quickly when the number of attributes of the instances grows, but there are some noise reduction techniques that may help to improve the accuracy [64] [65].

The K-nearest neighbour classifier was the first of the three classifiers implemented in this work.

This classification method was the fastest and simplest classification algorithm implemented and was based in two major steps:

1. Find the k instances in the training set that are closest to the instance that we are testing, according to some metric.
2. Majority vote of the k nearest instances to the test instance to determine the test class of the test instance.

The distance can be calculated by different formulas. For continuous variables, the distance calculated can be Euclidean, Manhattan or Minkowski, but if the variables are categorical, the distance to be calculated is usually the Hamming distance.

Euclidean Distance

$$\sqrt{\sum_{i=1}^K (x_i - y_i)^2} \quad (4.11)$$

Manhattan Distance

$$\sum_{i=1}^K |x_i - y_i| \quad (4.12)$$

Minkowski Distance

$$\left(\sum_{i=1}^K (|x_i - y_i|^q) \right)^{\frac{1}{q}} \quad (4.13)$$

Hamming Distance

$$D_H = \sum_{i=1}^K |x_i - y_i| \quad (4.14)$$

$$x = y \Rightarrow D = 0 \quad x \neq y \Rightarrow D = 1$$

In this work it was used the Euclidean distance to calculate the distances between the train instances and the test instances.

An example of how the algorithm works is shown in Figure 4.8. In the figure it is possible to see the test instance, as well as the train instances of this example. As the k is equal to 3, it was determined the three closest train instances to the test instance. Those instances are marked in the image, and it is also possible to see that, from that three instances, we have two belonging to class 1 and just one instance belonging to class 2, which means that the test instance will be classified as belonging to class 1.

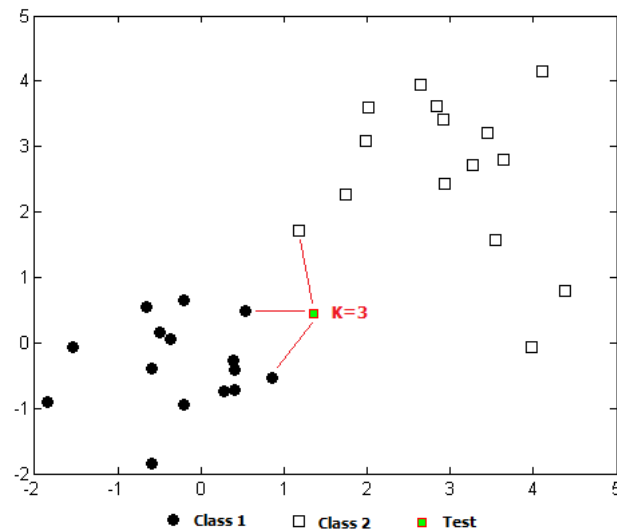


Figure 4.8: Example of KNN

The choice of k is the real challenging task in KNN algorithm. Choosing a higher k makes the algorithm less sensitive to noise and less precise but choosing a lower k makes the algorithm less stable, with higher variance. It was opted to test k from 1 to 13 that was when the TN rate started to get really low.

Given the original purpose of this work - not miss any malignant cases - two variations of the Simple KNN Algorithm were introduced:

1. Malignant Neighbour Algorithm
2. Asymmetric Distances KNN Algorithm

Those two variations are explained in the next subsections 4.4.1.1 and 4.4.1.2.

4.4.1.1 Malignant Neighbour Algorithm

The Malignant Neighbour algorithm is also a very simple algorithm that privileges the malignant class, which means that privileges the classification of malignant mammograms, as the purpose of this work is not to miss any malignant cases.

This variation of the KNN algorithm is based in two major steps:

1. Find the k instances in the training set that are closest to the instance that we are testing, according to some metric.
2. Determines if in the k closest instances, at least one has the malignant class.
 - If at least one of the neighbours is malignant, the tested class is classified as belonging to the malignant class.

- If all the neighbours have the benignant class, the tested class considered to be benignant.

As it is possible to see, this variation of the algorithm is quite similar to the Simple KNN algorithm.

4.4.1.2 Asymmetric Distances KNN Algorithm

The variation of the KNN algorithm using asymmetric distances was more complicated to implement than the other variation.

Such as the malignant neighbour variation of the KNN algorithm, this variation is as well based in the simple KNN algorithm. However, unlike the simple KNN algorithm where the distances were calculated in the same way for both classes, this variation of the algorithm stands out precisely because it calculates the distances in different ways according to the classes to which it belongs. Although the formula used to calculate the distances is the same, the Euclidean distance, for the malignant class, it was introduced a new parameter α to approach the malignant instances to the test instance.

The parameter α is a multiplicative factor applied to the Hamming distance for the malignant classes. As the intention of the α is to approach the malignant instances, the value of α is determined experimentally and varies between 0 and 0,9 if we are trying to approach the instance and, if trying to alienate, the value of α should be higher than 1. Having an malignant instance, where the Hamming distance was multiplied by an α of 0,5 means that those instance will be at $\frac{1}{5}$ of the original Hamming distance without the parameter α .

4.4.2 Support Vector Machines

As it was explained in 3.2.5, the Support Vector Machines classification method is based on the concept of decision planes that define decision boundaries. A decision plane separates between a set of objects having different class memberships. There are two main types of SVM classification methods: binary SVM classifiers and multi-class SVM classifiers.

In this work it was used a binary SVM classifier in order to compare the results obtained with the KNN classifier and the Random Forests classifier that was implemented afterwards.

The SVM classifier was chosen because it finds the optimal separation hyperplane and maximizes the margin between classes. That fact makes the classification more accurate. The SVM classifier usually presents better results than other classifiers and is more robust to the problem of overfitting.

In terms of limitations, one of the main limitations is make the correct choice of the kernel function, as well as its parameters, when using a non-linear SVM, but the complexity, transparency and requirements of memory and CPU time are also big limitations in the use of this classification method [10].

To test this algorithm two functions from the Statistics Toolbox of MatLab were used: *svmtrain* and *svmclassify*.

The function *svmtrain* returns a structure, *SVMStruct*, containing information about the trained SVM classifier [66]. In the simplest form, only the train features and the train classes are passed to the function as arguments, but to improve the results of the classification of the test features additional parameters should be considered:

- **autoscale:** this parameter specifies, when set to true, if the *svmtrain* function automatically centres the data points at their mean, and scales them to have unit standard deviation, before training.
- **boxconstraint:** specifies the value of the box constraint *C* for the soft margin.
- **kernelcachelimit:** this parameter specifies the size of the kernel matrix cache for the Sequential Minimal Optimization (SMO) training method. The default value of the kernel matrix is 5000.
- **kernel function:** this parameter specifies the type of kernel function used by the function to map the training data into the kernel space. The kernel function can be:
 - **linear:** linear kernel, which means the dot product
 - **quadratic:** quadratic kernel
 - **polynomial:** polynomial kernel
 - **rbf:** Gaussian Radial Basis Function kernel
 - **mlp:** Multilayer Perceptron kernel
 - **@kfun:** Function handle to a kernel function. It can be a function designed by the user

The default kernel function is the linear.

- **kktviolationlevel:** specifies the number of variables that are allowed to violate the Karush-Kuhn-Tucker (KKT) conditions for the SMO training method. The default value of this parameter is 0.
- **method:** Specifies the method used when separating the hyperplane. The method could be:
 - **QP:** Quadratic Programming. The classifier is a 2-norm soft-margin SVM.
 - **SMO:** Sequential Minimal Optimization.
 - **LS:** Least squares.

By default, the method used by the function is the SMO method.

- **mlp params:** Used when the kernel function is the mlp function. The default value is [1 -1].
- **options:** Additional parameters can be specified in this field. When the used method is SMO, the options parameter should be created using the *statset* function. If the method is the QP method, the options parameter should be created using the *optimset* function.
- **polyorder:** Used when the kernel function used is the polynomial function. The default value of the parameter is 3.

- **rbf sigma:** Used when the kernel function used is the rbf function. The default value of sigma is 1.
- **showplot:** The parameter showplot, when set to true, creates a plot of the class data of the training features and separating line. It only creates the plot if the data has exactly two features. The parameter is, by default, set to false.
- **tolkkt:** This parameter specifies the tolerance with which the KKT conditions are checked for the SMO training method. The default value of this parameter is $1e^{-3}$.

The structure SVMStruct returned by the function, contains information about the SVM classifier that has been trained by the function *svmtrain*.

The structure contains information about:

- **SupportVectors:** Matrix of data points with each row corresponding to a support vector in the normalized data space.
- **Alpha:** Vector of weights for the support vectors.
- **Bias:** Intercept of the hyperplane that separates the two groups in the normalized data space.
- **KernelFunction:** The kernel function used to map the training data.
- **KernelFunctionArgs:** Cell array of any additional arguments required by the kernel function.
- **GroupNames:** Specifies the group identifiers for the support vectors.
- **SupportVectorIndices:** Vector of indices that specify the rows in the training data, that were selected as support vectors after the data was normalized, according to the parameter autoscale set to true.
- **ScaleData:** Field containing normalization factors. Only contains data when the parameter autoscale is set to true, in that case, the structure contains two fields:
 - **shift:** Row vector of values where each value is the negative of the mean across an observation in the training data.
 - **scaleFactor:** Row vector of values where each value is 1 divided by the standard deviation of an observation in the training data.
- **FigureHandles:** Vector of figure handles created by *svmtrain* when using the parameter showplot.

The function *svmclassify* uses the SVMStruct, produced by the *svmtrain* function, to predict the class of the test instances. The arguments passed to the function are the SVMStruct and the test features that we want to have its classes predicted. The function classifies each row of the test features data using the information contained in the SVMStruct [67].

Similarly to the train instances, each row of the test instances corresponds to an observation, and each column corresponds to the features of an instance. Due to this fact, the test instances must have the same number of columns as the train instances, which means that the number of features should be the same for both of them.

Additionally, a parameter showplot can be assigned to the function in order to plot the test features, similarly to what happens in the *svmtrain* function where, when the showplot parameter is set to true, a plot with the classes of the training features is produced. This plot only appears when the data used to the production of the plot is two-dimensional.

In this work, some of the parameters of the *svmtrain* function have been changed from the default values. The kernel function that was chosen to be used was the mlp function, defined by the Formula 4.15.

$$K = \tanh(P_1 \times U \times V' + P_2), \quad (4.15)$$

where $P_1 > 0$ and $P_2 < 0$

It was chosen to be used the mlp kernel function after the realization of some tests, using the several kernel functions available. As the mlp kernel function obtained better results in this preliminary tests, it was opted for the use of this kernel function in all of the subsequent tests.

The parameters *mlp params* ($[P_1 P_2]$) and *boxconstraint* have been set to vary from 0.1 to 5 with a 0.2 step, in order to find the optimal parameters values, that are the values that produce a higher TP and TN rates when classifying the test instances. Once the optimal value was found, it was refined in one value with a 0.01 step, centred in that optimal point, in order to determine if a better result could be found.

4.4.3 Random Forests

Finally, it was tested the Random Forests algorithm for classification.

As it was said in 3.2.3, RF is a bootstrapping method applied to classification trees. It grows several classification trees and each classification tree is a complete tree with leaves that consist of some pre-selected number of observations. The prediction of the classes through RF method is made by aggregating the predictions of the ensemble of trees.

This classification method was chosen to be part of this work because it is a classification method with a simple training phase and fast in the classification task although the time to run the algorithm depends on the number of trees used to grow the forest [68]. The randomness associated with the RF algorithm provides notable robustness to noise, outliers, and over-fitting, when compared with only one. This randomness is also very important in terms of the computational efficiency because of the way the trees are built.

When building a single tree, the model may select a random subset of the training dataset, not all the dataset, and at each node in the process of building the tree, only a small part of all variables are considered, which makes a considerable reduction in the computational demands [69]. Although it is an accurate method, the predictions made by RF algorithm are difficult for humans to interpret and they are a bit inconsistent.

To test this algorithm two functions from the Statistics Toolbox of MatLab were used: *TreeBagger* and *predict* [70] [71].

The function *TreeBagger* creates an ensemble of decision trees for the prediction of instances. To be able to operate the function, in its simplest form, the train features and the train classes, as well as the number of trees in the forest should be arguments of the function. In order to improve the results of the prediction of the test features, additional parameters should be considered in the function:

- **FBoot:** Fraction of input data to sample with replacement from the input data for growing each new tree. The default value is 1.
- **OOBPred:** When set to on, stores information about what observations are out of bag for each tree. It can be used to compute the predicted class probabilities for each tree in the ensemble. Default is off.
- **OOBVarImp:** When set to on, stores out-of-bag estimates of feature importance in the ensemble. Default is off.
- **Method:** Classification or Regression method for the random forest.
- **NVarToSample:** Number of variables to randomly select for each decision split. The default value is the square root of the number of features for each instance. The default is the square root of the number of variables for classification and one third of the number of variables for regression.
- **NPrint:** Number of grown trees after which *TreeBagger* displays a diagnostic message showing training progress. Default is no diagnostic messages.
- **MinLeaf:** Minimum number of observations per tree leaf. The default value is 1 for classification and 5 for regression.
- **Options:** Structure that specifies additional options that are used to control the computation when growing the ensemble of decision trees. There are three different options:
 - **UseParallel:** When set to true computes decision trees drawn on separate bootstrap replicates in parallel. Default is false.
 - **UseSubstreams:** When set to true selects each bootstrap replicate using a separate substream of the random number generator. Default is false.
 - **Streams:** A *RandStream* object or cell array of such objects.

In this work, the arguments chosen to be altered from the default values were, along with the number of trees, the arguments: *MinLeaf*, *NVarToSample* and *method*.

The *method* used was classification method, the *MinLeaf* used was 4 and 8 and, finally, the *NVarToSample* used was 2 and 3. In terms of the number of trees used to grow the forest, in the beginning, 3 trees were used. That number was gradually increased till have a forest with 500 trees.

Additionally to these arguments of the function, another arguments may be considered to improve the classification task.

The arguments of the function *classregtree*, with exception of 'minparent' argument, can also be used in this function. The function *classregtree* has a lot of arguments that can be used to improve the results. In this work, it was used only the argument *weights*, a vector of observation weights in order to have different weights for the different classes.

The function *predict* computes predicted classes of the trained ensemble produced by the function *treebagger* for the test instances. Usually the arguments passed to the function are just the structure obtained with the *treebagger* function and the test features but other arguments can be used in the function. Arguments such as:

- **trees:** Array of tree indices to use for the prediction of the classes. Default is all.
- **treeweights:** Array NTrees weights for weighting votes from the specified trees.
- **useifort:** Matrix indicating which trees are used to make predictions for each observation. By default all trees are used.

The function *predict* was used in its simplest form, using as arguments only the structure and the test features, which means that all the trees were used with the same weight in the prediction of test classes.

4.5 Evaluation

The evaluation of the accuracy, sensitivity and specificity of the classification was determined using the False Positive, False Negative, True Positive (TP) and True Negative (TN) rates.

- The FP is defined as the number of normal mammograms classified as abnormal.
- The FN is defined as the number of abnormal mammograms classified as normal.
- The TP is defined as the number of abnormal mammograms classified as abnormal.
- The TN is defined as the number of a normal mammograms classified as normal.

The sensitivity is calculated using the TP and FN rates and is defined as the rate of the true positives in all the mammograms with cancer. It can be determined using the following formula:

$$Sensitivity = \frac{TP}{TP + FN} \quad (4.16)$$

The specificity is calculated using the TN and FP rates and is defined as the rate of the true negatives in all the mammograms without cancer. It can be determined using the following formula:

$$Specificity = \frac{TN}{TN + FP} \quad (4.17)$$

The accuracy is calculated using the four rates: TP, TN, FP and FN, and is defined as the rate where it is measured the ability to correctly classify mammograms with or without cancer.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.18)$$

The Sensitivity and Specificity are referred, in chapter 5, as TP rate and TN rate.

In the evaluation of the experimental results, it was considered that a good classification of mammograms was accomplished when it was possible to obtain a TP rate of 90% minimum and a TN rate of 30% minimum. These minimum values resulted of conversations with radiologists where it was considered that having a TP rate of 90% and a TN rate of 30% allows to have a significative gain in terms of time and an acceptable performance. This minimum specifications were considered, having an higher TP rate, because it was more important to correctly classify the abnormal mammograms than the normal ones, as we want to automate the classification of abnormal mammograms.

Chapter 5

Results

This chapter presents the results of the different tested scenarios occurred during this work. The tested scenarios were:

1. Extraction of GLCM and LBP features with previous density classification
2. Extraction of features with previous density classification
3. Extraction of GLCM and LBP features without previous density classification

An important point to retain in the beginning of this chapter is that the results that we are longing to obtain have some minimum specifications. We expect to get a minimum TP rate of 90% and a minimum TN rate of 30%, which means that we should have an accuracy of 60% minimum.

5.1 Extraction of GLCM and LBP features with previous density classification

In this first scenario, after the density classification by experts, the mammograms were separated into two sets: the dense mammograms set and the fatty mammograms set. For each one of these sets, GLCM and LBP features were extracted separately from the mammograms, and those features were then used separately to classify the mammograms into benign mammograms or malign mammograms, using the three mentioned classifiers.

5.1.1 K-Nearest Neighbour Classifier

In Tables 5.1 and 5.2 are presented the results of the KNN classifier, for each one of the different types and versions of features extracted. The TP and TN rates of the classification of dense mammograms, are presented in Table 5.1, and the results of the fatty mammograms in Table 5.2.

The presented results for both tables correspond to the best results achieved with the KNN classifier using not only the simple KNN classifier, as well as the two variations presented. In the tables it is mentioned to which version the best result, for each type of features, belongs to.

From the analysis of the Table 5.1, it is possible to conclude that, when analysing the effect of the extraction of different features from dense mammograms, the GLCM features achieved a TP rate of 100%, it did not miss any malignant mammogram, although it only achieved a TN rate of 23%, which is a low rate. Comparing the results of the GLCM features with the results of $LBP_{8,1}$ features using only the mean and standard deviation as features, that was the set of features in all versions of the LBP features sets, that obtained a higher TN rate, it can be deduced that $LBP_{8,1}$ achieved a better result, if we take into account the accuracy of the type of features.

Table 5.1: Results of the KNN Classifier for GLCM and LBP Features - Dense Tissue

Type of Features	Number of Features	TP rate	TN rate	Accuracy	Type KNN
GLCM		100%	23,25%	61.63%	Asymmetric Distances
Simple LBP	2 Features	95%	11,62%	53,31%	Malignant Neighbour
Simple LBP	5 Features	100%	9,30%	54.65%	Malignant Neighbour
$LBP_{8,1}$	2 Features	95%	41,86%	68.43%	Asymmetric Distances
$LBP_{8,1}$	5 Features	90%	25,58%	57.79%	Asymmetric Distances
$LBP_{16,2}$	2 Features	95%	20,93%	57.97%	Asymmetric Distances
$LBP_{16,2}$	5 Features	*	*	*	

Although it misses one malignant mammogram, as it has a TN rate that is almost the double of the TN rate using the GLCM features, the $LBP_{8,1}$, has better results in terms of the accuracy.

Table 5.2: Results of the KNN Classifier for GLCM and LBP Features - Fatty Tissue

Type of Features	Number of Features	TP rate	TN rate	Accuracy	Type KNN
GLCM		97,56%	12,12%	54,84%	Asymmetric Distances
Simple LBP	2 Features	97,56%	17,17%	57,37%	Malignant Neighbour
Simple LBP	5 Features	97,56%	16,16%	56,86%	Malignant Neighbour
$LBP_{8,1}$	2 Features	92,68%	32,32%	62,50%	Malignant Neighbour
$LBP_{8,1}$	5 Features	92,68%	32,32%	62,50%	Malignant Neighbour
$LBP_{16,2}$	2 Features	100%	8,08%	54,04%	Asymmetric Distances
$LBP_{16,2}$	5 Features	*	*	*	

From the analysis of the Table 5.2, it is possible to say that a better result is achieved when using, in general, LBP features. For a more particular analysis between the LBP features, the version of LBP features that obtained better results was the two multi-resolution LBP features using 8 neighbours with a radius of 1.

The last rows of each one of the tables have the symbol * because, for the mentioned type of features $LBP_{16,2}$ using 5 features, the classifier did not provide usable results, as one of the five features extracted had *NaN* values.

For the overall analysis of the KNN classifier, which can be seen in Figure 5.1, it is possible to conclude that, in general terms, the KNN classifier works better in dense mammograms, and it is harder to classify the mammograms constituted mostly by fatty breast tissue. The TN rate obtained for the classification through features extracted from fatty mammograms decreased when compared with the ones extracted from dense mammograms although the TP rate does not change a lot.

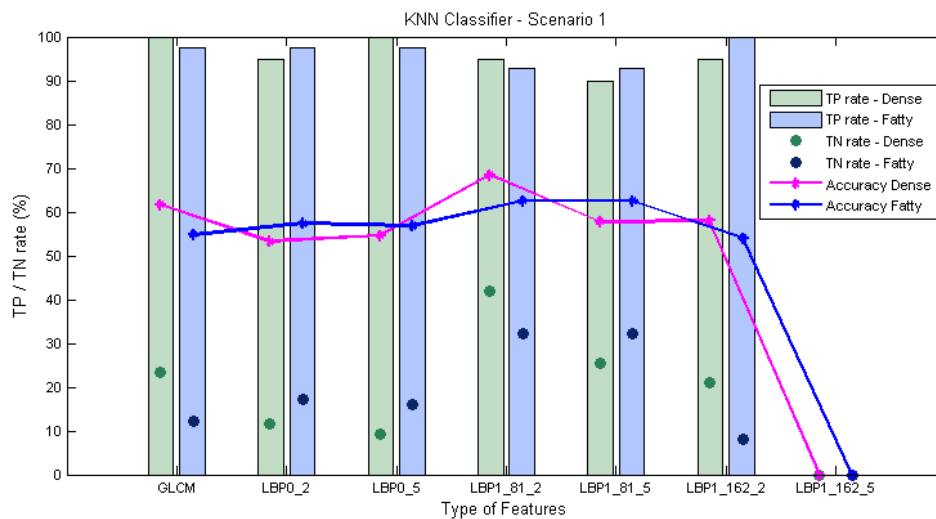


Figure 5.1: Results of the KNN classifier

In general, the LBP features obtained better results than the GLCM features, and the features that obtained better results for both dense and fatty mammograms were the multi-resolution LBP features - $LBP_{8,1}$ - using the mean and standard deviation features. In the analysis of all the features extracted from both dense and fatty mammograms, the type of features where only two of the five features extracted were used, accomplished similar or better results.

5.1.2 Support Vector Machines Classifier

In Tables 5.3 and 5.4 are presented, for each one of the different features extracted, the results of the SVM classifier. The TP and TN rates of the classification of dense mammograms, are presented in Table 5.3, and the results of the fatty mammograms, in Table 5.4.

Similarly to what happened for the KNN classifier, the rows having the symbol *, for the type of features $LBP_{8,1}$ and $LBP_{16,2}$ using 5 features, the classifier did not provide usable results.

Through the analysis of the Table 5.3, it is possible to see that LBP features have better results for both rates, when compared with GLCM features. The $LBP_{8,1}$ using only the mean and the standard deviation features, were the ones that achieved better results.

Table 5.3: Results of the SVM Classifier for GLCM and LBP Features - Dense Tissue

Type of Features	Number of Features	TP rate	TN rate	Accuracy
GLCM		90%	44,19%	67,10%
Simple LBP	2 Features	95%	86,05%	90,53%
Simple LBP	5 Features	95%	41,86%	68,43%
$LBP_{8,1}$	2 Features	100%	90,69%	95,35%
$LBP_{8,1}$	5 Features	*	*	*
$LBP_{16,2}$	2 Features	100%	86,05%	93,03%
$LBP_{16,2}$	5 Features	*	*	*

As said in the beginning of this chapter, when this work started to be done, our goal was to obtain TP rates around 90% and TN rates around 30%. Managing to achieve a TP rate of 100% along with a TN rate of 90,69% is considered an extremely good result. This means that using this combination of classifier plus features, allowed it to not miss the correct classification of any malignant mammogram and only misclassify 9 benign mammograms. This particular result is extremely satisfying when designing a pre-CAD system.

Table 5.4: Results of the SVM Classifier for GLCM and LBP Features - Fatty Tissue

Type of Features	Number of Features	TP rate	TN rate	Accuracy
GLCM		90,24%	39,39%	64,82%
		92,68%	10,10%	51,39%
Simple LBP	2 Features	100%	75,76%	87,88%
Simple LBP	5 Features	92,68%	45,45%	69,07%
$LBP_{8,1}$	2 Features	97,56%	77,78%	87,67%
$LBP_{8,1}$	5 Features	90,24%	55,56%	72,90%
$LBP_{16,2}$	2 Features	100%	80,81%	90,41%
$LBP_{16,2}$	5 Features	*	*	*

For fatty mammograms, through the Table 5.4, it is possible to see the results of the SVM classifier. The results were also very positive, mainly for LBP features that obtained much better results than the GLCM features. For the LBP features, $LBP_{16,2}$ was the type of LBP features that obtained better results, with an accuracy of 90,41%.

For the overall analysis it was possible to observe several things.

In accuracy terms, the dense mammograms achieved better results, mostly because they were higher than the results for the fatty mammograms, although none of them were stable results for the classifier in general.

Similarly to the KNN classifier, the features that obtained better global results were the LBP descriptors, $LBP_{8,1}$ using the mean and standard deviation features, although that for fatty mammograms, the features sets that obtained better results were the multi-resolution LBP version, using only two features.

In Figure 5.2 are represented the results for both dense and fatty mammograms.

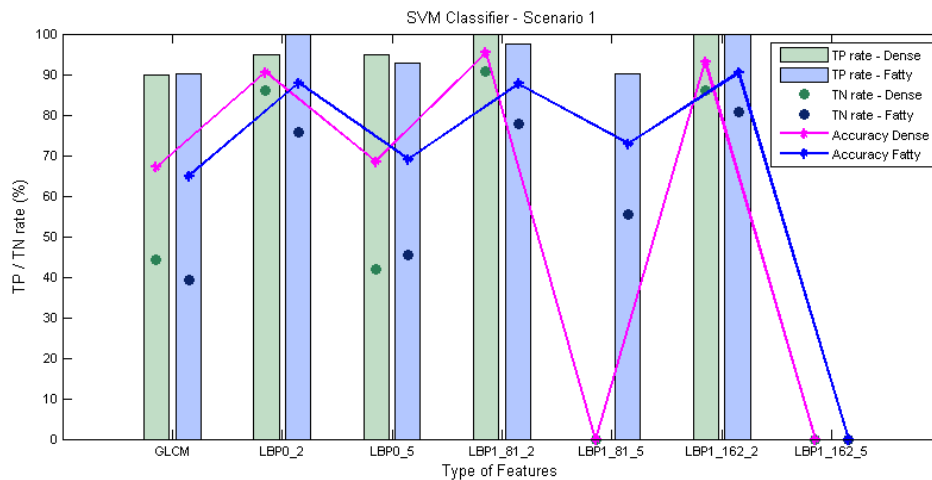


Figure 5.2: Results of the SVM classifier

It is also possible to conclude that, in general terms, the SVM classifier works better in dense mammograms, similarly to the KNN classifier. The TN rate obtained for the classification through features extracted from fatty mammograms was above the TN rate for the same version of feature extracted from the dense mammograms, except for Simple LBP features using the mean and standard deviation.

5.1.3 Random Forests Classifier

In Tables 5.5 and 5.6 are presented the results of the RF classifier, for each one of the different features extracted. The TP and TN rates of the classification of dense mammograms, are presented in Table 5.5, and the results of the fatty mammograms, in Table 5.6.

Table 5.5: Results of the RF Classifier for GLCM and LBP Features - Dense Tissue

Type of Features	Number of Features	TP rate	TN rate	Accuracy
GLCM		91%	14,45%	52,73%
Simple LBP	2 Features	99%	6,74%	52,87%
Simple LBP	5 Features	95,5%	7,21%	51,36%
$LBP_{8,1}$	2 Features	97%	9,77%	53,39%
$LBP_{8,1}$	5 Features	96%	5,814%	50,91%
$LBP_{16,2}$	2 Features	97%	0%	48,50%
$LBP_{16,2}$	5 Features	95,5%	0,93%	48,22%

Through the Table 5.5 it is possible to see that the features that had better results for this classifier were the $LBP_{8,1}$ using only the mean and standard deviation features. Although this is the version of LBP features that obtained better results, when compared with the other versions

for this type of features, the others had very similar results: good TP rates but very low TN rates, which makes the overall result of the classifier far below than desired.

Table 5.6: Results of the RF Classifier for GLCM and LBP Features - Fatty Tissue

Type of Features	Number of Features	TP rate	TN rate	Accuracy
GLCM		92,68%	10,51%	51,60%
Simple LBP	2 Features	89,51%	28,28%	58,90%
Simple LBP	5 Features	81,95%	28,08%	55,02%
$LBP_{8,1}$	2 Features	85,37%	13,23%	49,30%
$LBP_{8,1}$	5 Features	80,98%	24,24%	52,61%
$LBP_{16,2}$	2 Features	86,83%	18,49%	52,66%
$LBP_{16,2}$	5 Features	86,34%	18,99%	52,67%

When analysing the RF classifier for fatty tissues, through the Table 5.6, it is possible to see that the TN rates are substantially higher than the TN rates for dense tissues but the TP rates decreased. The set of features that had better results with RF classifier were the Simple LBP using only two features.

Taking into account the accuracy of the classifier for the different features, visible in Figure 5.3, in the overall analysis, the accuracy is higher for fatty mammograms than for dense mammograms but the values are not between the acceptable values: TP higher than 90% and TN higher than 30%. The version that obtained better results for both dense and fatty mammograms is the Simple LBP features version, using only two of the five possible features. In this version, the TP and TN rates were almost between the minimum values.

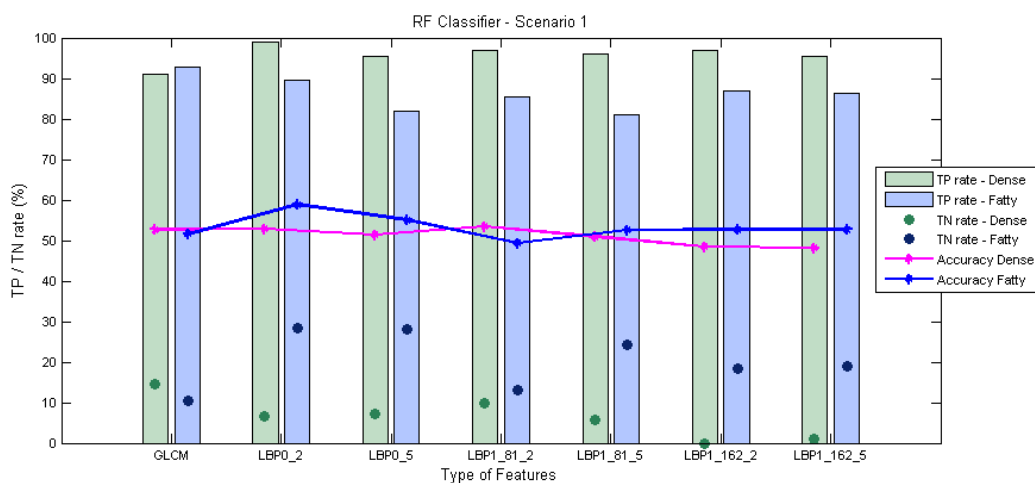


Figure 5.3: Results of the RF classifier

This classifier was the one where the results were lower. Although we have achieved a TP rate higher than the 90%, considered the minimum acceptable, the TN rates were extremely low. Having this combined results means that almost all mammograms were classified as malignant,

which is not a good practice in a pre-CAD system. Although it is true that our main objective is to automate the malignant classification without missing malignant mammograms, and not properly automate the benign cases, saying that all mammograms are malignant is not considered a good result. Another drawback of the RF classifier was the time of running of the classifier. It was a very low classifier, which could be compensated if the achieved results were extremely good results, which did not happen.

5.1.4 Conclusion Scenario 1

The results obtained with the SVM classifier were much better than the results of the two other classifiers. The rate of true positives is, in a global perspective, higher than in the other classifiers, as well as the rate of true negatives. With the SVM classifier it was possible to achieve TN rates much higher than was initially expected. Those rates with the SVM classifier were the double, sometimes the triple of the TN rates obtained when using the other classifiers. Another factor that influenced the choice of the SVM as the best classifier was the time for the algorithm to run. The SVM classifier, although not the fastest classifier, was fast in the classification, compared with the RF classifier. This scenario presents the dense tissue type as the tissue where the results are more satisfactory. Although there are also good results in fatty mammograms, the dense ones presented, in general, higher results for this scenario.

In the comparison of the performance of GLCM and LBP features, the LBP features came out victorious. In a more particular comparison between the different types of LBP features extracted from the mammograms, there were different results according to the different types of classifiers used. Even though, the Multi-Resolution LBP features were the features that achieved better results for two of the classifiers: the KNN and the SVM classifiers. For the RF classifier, the type of features that obtained better results were the Simple LBP features.

Another point to take into account in the features performance in this scenario is that, between all the types of LBP features used to test this scenario, the ones that obtained better results between the Simple LBP features and the Multi-Resolution LBP features, were always the versions where it was used only two, and not five, features. Through this it can be concluded that it is not important to extract a lot of features, but the right features to use in the classification.

An important part of this work was to understand if the conclusions made by Elshinawy *et al* in [46] [27] were considered also true for our dataset and our work. The results obtained in this work were not quite similar to Elshinawy *et al* work. Elshinawy concluded in her work that the GLCM features performed better in fatty breasts and LBP features in dense breasts. As it is possible to see from the tables presented above, for all three classifiers the GLCM features achieved better results in mammograms with dense tissue, which is the reverse. On the other hand, for LBP features, for the KNN classifier and the SVM classifier it was possible to corroborate with our work that, as Elshinawy *et al* concluded, the LBP features performed better in dense tissues, although in our case, using the RF classifier, we achieved better results in fatty tissues.

5.2 Extraction of features with previous density classification

Similarly to the first presented scenario, in this scenario the mammograms were also classified according to the density of the mammograms. From each one of the sets constituted after the classification by experts, unlike what happened in the first scenario where the features were extracted and classified separately, in this new scenario, GLCM and LBP features were extracted from each mammogram and then used together in the classification.

Not all versions of the LBP features were extracted along with the GLCM features. As it was concluded in 5.1 for the first scenario, the classifier that achieved a better global result was the SVM classifier, thus, in order to determine which versions of the LBP features should be extracted, along with the GLCM features, it was determined, in scenario 1, which LBP features had a better result for dense mammograms with the SVM classifier, as well as for fatty mammograms, to use that information in this new scenario.

For dense mammograms, the features that obtained a better result for the SVM classifier were the Multi-resolution LBP features with $P = 8$ and $R = 1$ ($LBP_{8,1}$), using only two features. For fatty mammograms, the features that obtained better results for the SVM classifier were the Multi-resolution LBP features with $P = 16$ and $R = 2$ ($LBP_{16,2}$), using only two features.

5.2.1 Dense

Table 5.7 presents, for dense mammograms, the results of the classification using both KNN and SVM classifiers.

Table 5.7: Results for Dense tissue for KNN and SVM classifiers

Classifier	TP rate	TN rate	Accuracy
KNN	90%	51,26%	70,63%
	95%	39,53%	67,27%
	100%	11,63%	55,82%
SVM	90%	69,77%	79,89%
	95%	11,63%	53,32%
	100%	11,63%	55,82%

For dense mammograms, the classifiers had a very similar performance, if we think on the results in average accuracy terms. The average results of the accuracy for the best values of both classifiers were 64,58% and 63,01%, for the KNN classifier and the SVM classifier, respectively. It is not a very uneven result, although the KNN classifier achieved the best average result, the SVM classifier obtained the best accuracy result for dense mammograms.

In the Figure 5.4 is shown the evolution, for dense mammograms, of the best results for both classifiers.

By analysing the Figure, it is easier to get a preview of the conclusions made through the Table 5.7, which one of the classifiers obtained a better performance for dense mammograms.

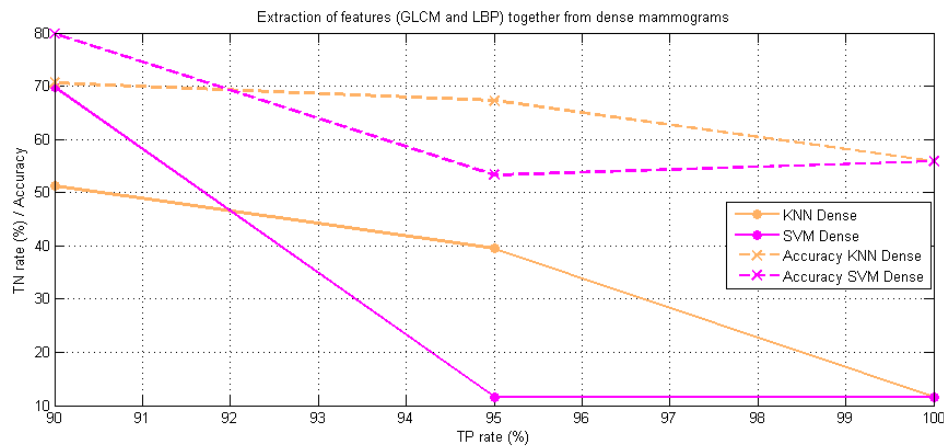


Figure 5.4: Results of the extraction of features together (GLCM and LBP) from dense mammograms with KNN and SVM classifiers

5.2.2 Fatty

Table 5.8 presents, for fatty mammograms, the results of the classification using both KNN and SVM classifiers.

Table 5.8: Results for Fatty tissue for KNN and SVM classifiers

Classifier	TP rate	TN rate	Accuracy
KNN	90,24%	16,16%	53,20%
	92,68%	13,13%	52,91%
	95,12%	11,11%	53,12%
	97,56%	10,10%	53,83%
	100%	9,09%	54,55%
SVM	90,24%	35,35%	62,80%
	92,68%	33,33%	63,01%
	95,12%	29,29%	62,21%
	97,56%	31,31%	64,44%

For fatty mammograms, the SVM classifier had a better performance when compared with the KNN classifier. Although the TP rates for both classifiers are the same, the TN rates for the SVM classifier are higher than the ones for the KNN classifier, and simultaneously they are above the minimum expected value, unlike the KNN classifier that had all TN rates for the best TP rates values much lower than it was expected.

In the Figure 5.5 is shown the evolution, for fatty mammograms, of the best results for both classifiers.

Through the figure, it is possible to conclude that all the best values for both classifiers, are more or less stable, for all the parameters analysed: TP rate, TN rate and accuracy. It is also easy to confirm that the SVM classifier is the one with most positive results for this tissue type.

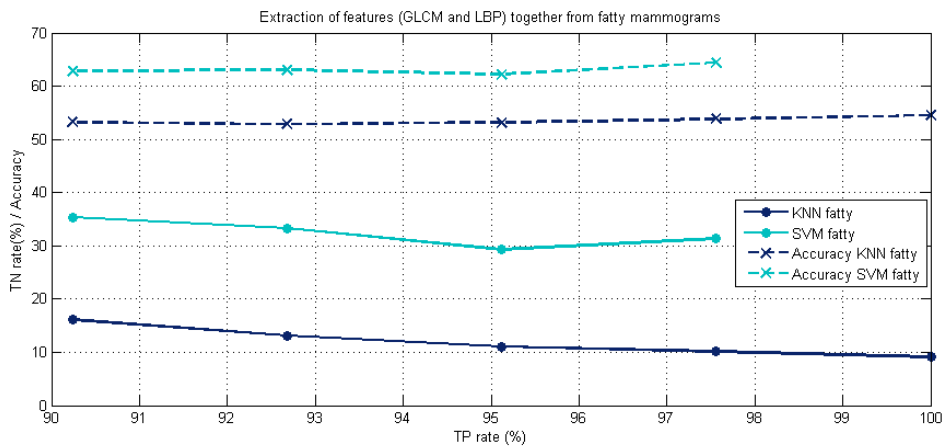


Figure 5.5: Results of the extraction of features together (GLCM and LBP) from fatty mammograms with KNN and SVM classifiers

5.2.3 Conclusion Scenario 2

As it was possible to see by both tables, the two classifiers performed better in dense tissues, achieving higher results for that type of tissue. Nevertheless, if we think on the results from a stability perspective, the classifiers were more stable when the extracted features belong to fatty mammograms.

In resemblance to what happened in the first scenario, in global terms, the classifier that achieved better results was the SVM classifier. If we consider the best result obtained for each classifier, for both dense and fatty mammograms. However, if we consider the minimum accuracy value, established in the beginning of this chapter, in the analysis of the best classifier for this scenario, we will have two different classifiers according to the tissue type. As the SVM classifier, is the unique classifier that has all of accuracy points over 60%, it is considered the best classifier for fatty tissues in this scenario. On the other hand, for dense tissues, there is not a classifier that has all accuracy points over the mark of 60%, thus, the classifier that has more points over 60%, is the best classifier for dense tissues.

By the Figures 5.4 and 5.5 it can be concluded a number of things:

1. The KNN classifier has a better performance in dense tissues than in fatty tissues
2. The SVM classifier has a better performance in fatty tissues
3. The SVM classifier has a better performance than the KNN classifier for both tissue types

Although it is not always true that the SVM classifier has better performance in fatty tissues than in dense tissues, which may be corroborated by the fact that the best result in this scenario is in dense tissues using the SVM classifier, in general terms, for dense breasts, the classifier is more stable and the accuracy is higher in the majority of the cases.

5.3 Extraction of GLCM and LBP Features without previous density classification

Unlike the other two tested scenarios, in this one the density classification was not taken into account, which means that there was only one data set. From this dataset, GLCM and LBP features were extracted separately, as in the first scenario, and then those features were used separately to classify the mammograms into benign mammograms or malign mammograms, using only the SVM classifier.

In this scenario, only the SVM classifier was tested because it was the one that, in general, obtained better results in both of the previous tested scenarios.

The Simple LBP and Multi-resolution LBP features extracted were the versions that used only two features: mean and standard deviation, for the same reason that we only tested the SVM classifier, the features were chosen because of the better results obtained in the first scenario.

Table 5.9 presents, for each one of the different features extracted, the results of the SVM classifier for this scenario.

Table 5.9: Results SVM Classifier for GLCM and LBP Features without previous density classification

Type of Features	Number of Features	TP rate	TN rate	Accuracy
GLCM		90,16%	78,17%	84,17%
		91,80%	75,65%	83,73%
		93,44%	76,06%	84,75%
		95,08%	73,24%	84,16%
		96,72%	74,65%	85,69%
		98,36%	40,85%	69,61%
		100%	45,07%	72,54%
Simple LBP	2 Features	80,33%	83,10%	81,72%
		81,97%	64,79%	73,38%
		83,61%	76,06%	79,84%
		85,25%	80,28%	82,77%
		86,89%	69,01%	77,95%
		88,52%	25,35%	56,94%
$LBP_{8,1}$	2 Features	80,33%	77,46%	78,90%
		81,97%	71,13%	76,55%
		83,61%	66,90%	75,26%
		85,25%	66,20%	75,73%
		86,89%	65,49%	76,19%
		88,52%	66,90%	77,71%
$LBP_{16,2}$	2 Features	80,33%	66,90%	73,62%
		81,97%	68,31%	75,14%
		83,61%	23,24%	53,43%
		85,25%	24,65%	54,95%
		86,89%	24,65%	55,77%

Through the analysis of the Table 5.9 it is possible to conclude which features obtained better results when the density of the tissue is not taken into account to the classification.

In this scenario, unlike what happened in the two other scenarios, the GLCM features achieved the best results in the classification of mammograms. Although the GLCM features obtained the best results for this scenario, where the density of the tissue was not taken into account when extracting features from the classifier, the other tested versions of features, also obtain very good results. For the LBP features, although the majority of the results has an accuracy values above the minimum specified value, the drawback with these other type of features is the values of the TP rates. When trying to automate the classification of malign mammograms, missing 15% to 20% of them it is a high value.

As it is possible to observe in Figure 5.6, for TP rates between 90% and approximately 97% the GLCM features using the SVM classifier obtained an accuracy around 85%, which is the equivalent to have TN rates around 75%.

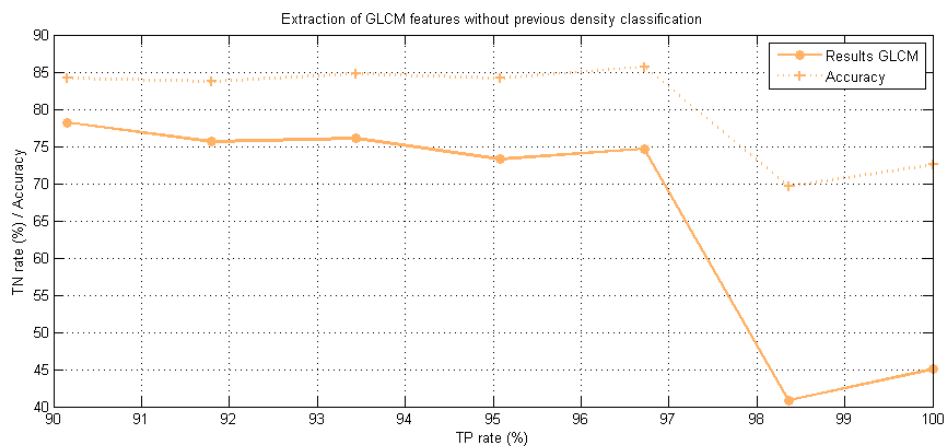


Figure 5.6: Results of the extraction of GLCM features without previous density classification

The results that we have obtained are extremely good results, taking into account that the density is not taken into consideration.

In Figure 5.7, it is possible to analyse the results for the Simple LBP features, with regard to the TP and TN rates, as well as the accuracy of those results.

Trough the analysis it is possible to conclude that Simple LBP features obtained a good accuracy for the best values achieved, although those values are not good in terms of TP rate. The best TP rates obtained were situated between 80% and 88%, which is a result below what we are expecting for, nevertheless the TN rates were extremely good and far higher than we originally

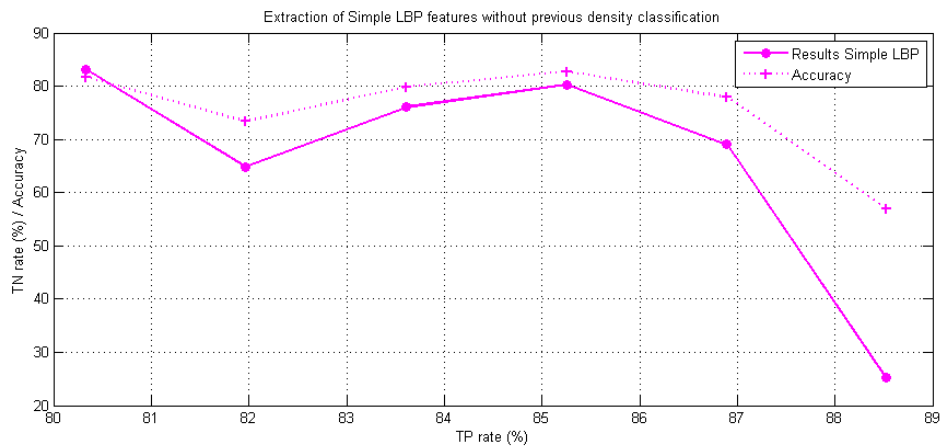


Figure 5.7: Results of the extraction of Simple LBP features without previous density classification

expected, apart the result with a TP rate of 88,52% where the TN rate is not around the other results but it is lower.

In Figure 5.8, it is possible to analyse the results for the Multi-Resolution LBP features, having $P = 8$ and $R = 1$.

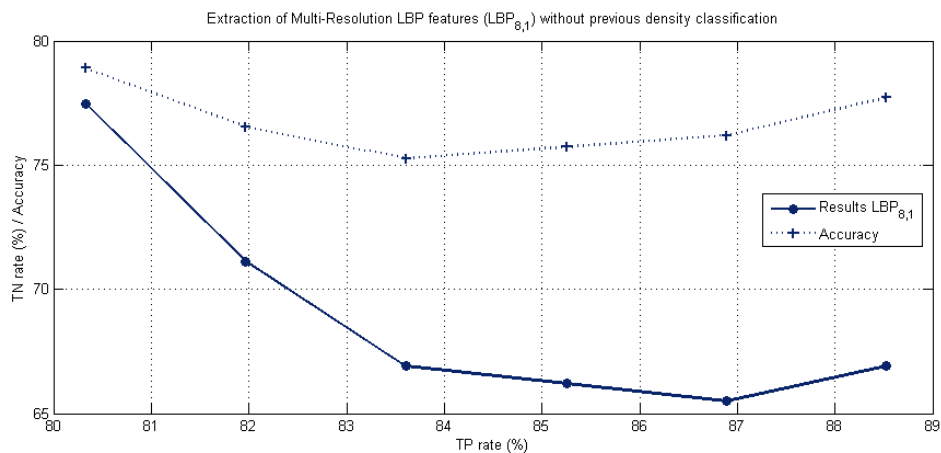


Figure 5.8: Results of the extraction of Multi-Resolution LBP features ($LBP_{8,1}$) without previous density classification

Although the results for this type of features were not as good as the results obtained for Simple LBP versions, the TN rate, for the best TP rate obtained, is much better than the one obtained for Simple LBP features versions. As the TP rates for both the types of features were the same and the accuracy were lower, it means that the TN rates were, in general, lower.

Through the analysis of Figure 5.9, it is possible to see more clearly the results for Multi-Resolution LBP features having $P = 16$ and $R = 2$.

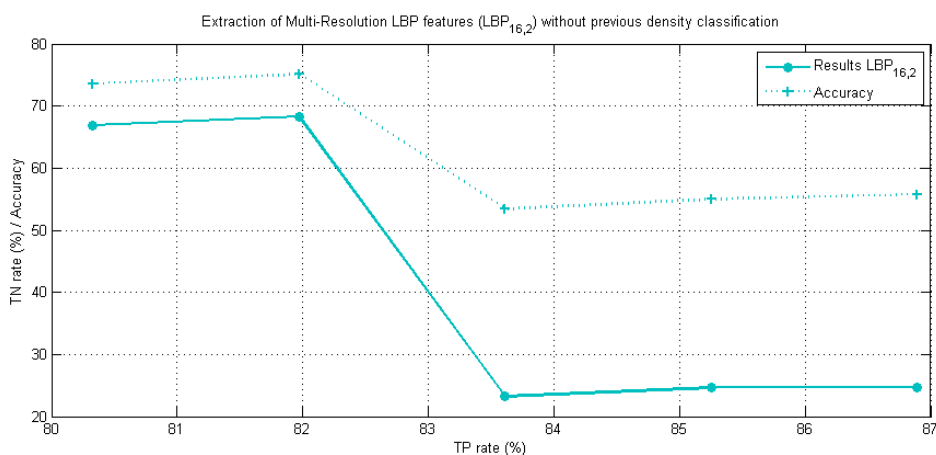


Figure 5.9: Results of the extraction of Multi-Resolution LBP features ($LBP_{16,2}$) without previous density classification

The Multi-Resolution LBP features having $P = 16$ and $R = 2$ were the ones that achieved worse results. Those results had a TP rate below 90%, that right from the start was not an encouraging result.

In the comparison of Figure 5.9 with the previous presented figures, it is immediately possible to see the despair of the results between the different versions of features, and the notable disadvantage of this type, having an accuracy, for all the values closer to the desirable TP rate of 90%, around 55%, which is a very low accuracy.

5.3.1 Conclusion Scenario 3

In this scenario, where the density of the breasts was not taken into account when extracting the GLCM and LBP features, it can be concluded that the GLCM features obtained the best results in the classification of mammograms.

Between the LBP features, the version of the Multi-Resolution features using $P = 8$, $R = 1$ and only the mean and standard deviation was the one that achieved better results, with an average accuracy of 76,72%, it has only a substantial difference to the Simple LBP features version that obtained an average accuracy of 75,43%.

The main drawback for the LBP features was the fact that, with the LBP features, the best TP rate achieved was 88,52%, a value under the desired result of 90%.

5.4 Conclusions

In the comparison of scenarios, it can be concluded that it is a major advantage to separate the type of features that are extracted and posteriorly used in the classification, not for the results of the TP rate but mostly for the results of the TN rate that are higher when the features used in the classification are not a group of GLCM and LBP features.

It can also be concluded by the tests with different scenarios, that separating the mammograms according to the density, for our database, it was not something so essential, although good results are obtained when that separation by density is made. The results obtained when this separation is not made, are also good and better than we expected. It was achieved, with the third scenario, where the density was not taken into account, very good results for GLCM features.

In the analysis of each scenario, it was possible to conclude that, in general, for the two scenarios where the density was taken into account, the classification of dense mammograms obtained better results than the classification of fatty mammograms.

The classifier that obtained better results was the SVM classifier, where it was possible to achieve better results than we were initially expecting. For some of the features, the classifier was able to achieve an accuracy rate of 95,35%, which is an extremely good rate. This result means that, in this specific case, we managed not to miss any malignant mammogram and only misclassify 4 benign mammograms. As our objective is to automate the classification of malignant mammograms, and as this system is intended to be a pre-CAD system to work as a “first look” in the classification of mammograms, misclassify 4 benign mammograms is not a very negative point.

The classifier that had a worse performance was the RF classifier, not only by the achieved results but also by the time that the classifier took to do only one classification, according to the defined parameters.

The KNN classifier, although not the best classifier, obtained good results, speciality for dense mammograms, and was also a fast classifier.

Finally, in terms of the features, it was concluded that the LBP features achieved better results than the GLCM features, in the majority of cases, for the first two scenarios. Although using several versions of LBP features, the Multi-Resolution LBP features were the version that obtained better results in more tests.

It was also possible to conclude that, when analysing the LBP features, both for Simple LBP features and Multi-Resolution features $LBP_{8,1}$ and $LBP_{16,2}$, the versions where only two features were extracted and used in the classification phase, achieved better results than the versions where it was used the five available features: mean, standard deviation, entropy, energy and skewness.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

During the research and the development of this work, it was possible to better understand the Breast Cancer problematic, as well as the pre-CAD systems, important systems that aim to help the radiologists in the pre-diagnostic of mammograms. As it is known, the radiologists and the medical staff have to analyse, each day, a large number of mammograms. Repetitive tasks tend to have an increased number of errors, thus, in order to help the medical staff reducing the number of errors which, in this specific case, means reducing the number of misclassifications of mammograms, the pre-CAD and CAD systems were created.

The main difference between pre-CAD and CAD systems is that one functions as a “first look” in the diagnosis of a mammogram, while the other functions as a “second look”. As our objective was to automate the classification of abnormal mammograms, this is, automate the classification of mammograms that we are not sure of its malignancy, and we want those mammograms to be classified later on by the medical staff, as we know misclassify a mammogram as benign and after all the mammogram turns to be malignant, without further analysis by an expert, at those cases that the system were not completely certain, will be extremely serious. Hereupon, a pre-CAD system was implemented instead of a CAD system.

The pre-CAD system implemented was based in four major steps: density classification by experts, pre-processing of the mammograms, features extraction and, finally, classification.

The mammograms were classified by experts into dense or fatty mammograms. After that initial classification, pre-processing was applied to the mammograms, in order to reduce the dimensionality and remove possible artifacts that might be present.

The two main steps in this system were the features extraction and the classification. The features extraction was one of the most important steps, as the classification depends on the extracted features to perform a good classification, achieving good TP and TN rates.

Two types of features were extracted: GLCM features and LBP features. In the GLCM type, ten features were extracted and used in the classification. On the other hand, the LBP features were a little more complicated to extract.

Two versions of LBP features were extracted: the Simple LBP features and a multi-resolution approach to grey-scale LBP features. With respect to the multi-resolution approach to grey-scale, two different arrangements were used: one using 8 neighbours and a radius of 1, and another using 16 neighbours and a radius of 2. For each one of the versions and arrangements of the LBP features, two sets of features were combined. In the first set, only two features were extracted and posteriorly used in the classification phase and, in the other set, five features were extracted and used in the classification.

In what concerns classification methods, three were used: K-nearest neighbours, Support Vector Machines and Random Forests.

For each one of the classifiers, several arguments were used in order to improve the classification of the mammograms. For the KNN classifier, two variations of the Simple KNN algorithm were proposed: the malignant neighbour KNN algorithm and the asymmetric distances KNN algorithm. The other two classifiers were used in the original form, specifying only additional arguments in the functions that were used to make them work properly.

After the implementation phase, it was started the tests phase. In this phase, three different scenarios were tested.

In the first scenario, GLCM and LBP features were extracted with previous density classification of the mammograms. The three classifiers were tested and very good results were obtained. The SVM classifier was the classifier that achieved better results in this scenario. The worst classifier was the RF classifier. It did not achieved good results, as the accuracy was always below the minimum value and the classifier was extremely slow, if the number of trees to grow the forest was big. The best results were obtained in dense mammograms, although there were also very good results for the SVM classifier in fatty mammograms. The LBP features, when compared with the GLCM features achieved better results.

This scenario has served to prove that separating the mammograms by tissue type is an added value to the system. The results proved to be very good, when using the SVM classifier, even better than it was initially expected. Nevertheless, one of the main points that we were longing to prove right, was not completely proven right. When this work started, one of the main purposes was to prove that the conclusions made Elshinawy *et al*, were also right for our dataset. In their work it was said that GLCM features performed better in mammograms of breast constituted by fatty tissues and LBP features performed better in mammograms of dense tissue. However it was always the LBP features that obtained better results for all classifiers. It was also proven that between only the GLCM features, the best results were always achieved for dense mammograms.

In the second scenario, the features were extracted and used together in the classification of mammograms, previously classified according to the density.

In this scenario, the best results were also obtained for dense mammograms, like what happened in the first scenario. However, if we think in the results in a stability perspective, the results obtained by the classifiers were more stable in fatty tissues.

In global terms, in this scenario, the classifier that obtained better results was the SVM classifier for both density types.

If we analyse each one of the classifiers, the KNN classifier achieved better results in dense tissues, although the SVM classifier obtained higher accuracy in fatty tissues.

With this scenario, we were able to conclude that having the features used together in the classification, it is not a good scenario. The results obtained were worse so, separating the features to use in the classification phase, which means using the GLCM and LBP features separately in the classification is an added value to the system, as better results were achieved in the first scenario than in this scenario.

Finally, in the third scenario, the density of the mammograms were not taken into account and the GLCM and LBP features were extracted from the full set of mammograms.

Unlike the other scenarios, in this third scenario, it was not the LBP features type that achieved better results. The GLCM features obtained better results, for our database, and those results were very close to the results of the SVM classifier for the first scenario.

The major conclusion that is made by the third scenario is that, it is not a major advantage to separate the mammograms by density. The obtained results did not decreased so much when they were not separated, and still the results were higher than the minimum expected value, having an average accuracy 10% higher than the minimum accuracy desired.

6.2 Future Work

Future work can be done to improve this Master's Dissertation.

To test the work done in other databases, seems to be one of the key points to understand if the good accuracy that was achieved for this database, is also achieved with other databases, providing this work the true possibility of an implementation in the medical environment. But, for this implementation to be possible, it should be in a more fast and efficient language, in order to accomplish a more efficient and fast system, it should be implemented in C, C++ or Java languages.

Testing new possibilities, as it is the example of new classifiers and features, as well as new kernel functions for the SVM classifier, could help realize if the options of classifiers and features to extract, made in this work were the right choices, or if we could have achieved even better results.

When trying to improve the system implemented, there are also a lot of options.

Starting with the use of both CC and MLO views of the mammography to extract the features, which may imply the elaboration of an algorithm to separate the pectoral muscle from the breast.

Another point that may be very interesting to implement to improve this work is the automate classification of the breast according to the density of the breast. In this work, such classification was done by experts. Although this automatic classification could represent many improvements in a more autonomous pre-CAD system, the automatic classification of the density, may raise some issues in the results. With this improvement, the results obtained in this work should decrease a little, which may compromise the viability of the pre-CAD.

References

- [1] Yale University. Introduction to cardiothoracic imaging, 2012. Available in http://www.yale.edu/imaging/notes/mammo_normal.html, last access in 04 November 2012.
- [2] Imaginis. Mammographic screening is key to the early detection of breast cancer, 2012. Available in <http://www.imaginis.com/breast-cancer-screening-prevention/mammographic-screening-is-key-to-the-early-detection-of-breast-cancer>, last access in 08 November 2012.
- [3] P. Samardar, E.S. de Paredes, M.M. Grimes, and J.D. Wilson. Focal asymmetric densities seen at mammography: Us and pathologic correlation. *Radiographics*, 22(1):19–33, 2002.
- [4] Steven B. Halls. Breast abnormalities typically discovered by mammogram, 2011. Available in <http://www.breast-cancer.ca/screening/mammogram-abnormalities.htm>, last access in 16 November 2012.
- [5] E.D. Pisano, E.B. Cole, B.M. Hemminger, M.J. Yaffe, S.R. Aylward, A.D.A. Maidment, R.E. Johnston, M.B. Williams, L.T. Niklason, E.F. Conant, et al. Image processing algorithms for digital mammography: A pictorial essay. *Radiographics*, 20(5):1479–1491, 2000.
- [6] Faculty of Medicine McGill. Interactive mammography analysis web tutorial, 2012. Available in <http://sprojects.mmi.mcgill.ca/mammography/>, last access in 1 December 2012.
- [7] Berg W.A. D’Orsi C.J., Bassett L.W. Breast imaging reporting and data system: ACR BI-RADS. *Breast Imaging Atlas, Reston, Va*, 2003.
- [8] R.L. Birdwell, D.M. Ikeda, K.F. O’Shaughnessy, and E.A. Sickles. Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection1. *Radiology*, 219(1):192–202, 2001.
- [9] A.S. Majid, E.S. de Paredes, R.D. Doherty, N.R. Sharma, and X. Salvador. Missed breast carcinoma: Pitfalls and pearls. *Radiographics*, 23(4):881–895, 2003.
- [10] M.Y. Elshinawy. *Pre-CAD normal mammogram detection algorithm based on tissue type*. PhD, Howard University, 2010.
- [11] Instituto Nacional de Estatística. Estatísticas no feminino: Ser mulher em portugal 2001-2011. Technical report, Instituto Nacional de Estatística, March 2012.
- [12] Liga Portuguesa Contra o Cancro. O cancro da mama, 2012. Available in <http://www.ligacontracancro.pt/gca/index.php?id=14/>, last access in 29 October de 2012.

- [13] Associação Laço. Laço - 10 anos. Technical report, Associação Laço, January 2010.
- [14] Portal da Oncologia Português. Os números do cancro da mama em portugal, 2012. Available in <http://www.pop.eu.com/portal/publico-geral/tipos-de-cancro/cancro-da-mama.html#numeros>, last access in 29 October 2012.
- [15] Monika Shinde. Computer aided diagnosis in digital mammography: Classification of mass and normal tissue, July 2003.
- [16] Trent D. Stephens Rod R. Seeley and Philip Tate. *Anatomia e Fisiologia*. LUSOCIÊNCIA, sixth edition, 2003.
- [17] International Atomic Energy Agency. Radiation protection of patients, 2012. Available in https://rpop.iaea.org/rpop/rpop/content/informationfor/healthprofessionals/1_radiology/mammography/mammography-technique.htm, last access in 04 November 2012.
- [18] Administração Regional de Saúde Norte. Circular normativa 03.11.2011. Technical report, Administração Regional de Saúde - Norte, November 2011.
- [19] National Cancer Institute. Mammograms, 2012. Available in <http://www.cancer.gov/cancertopics/factsheet/detection/mammograms>, last access in 04 November 2012.
- [20] Breastcancer.org. What are margins in relation to breast cancer?, 2012. Available in <http://www.breastcancer.org/questions/margins>, last access in 16 November 2012.
- [21] Pfarl G. Riedl C.C. and Helbich T.H. Breast imaging reporting and data system tutorial. Department of Radiology, University of Vienna, available in <http://www.birads.at>, last access in 6 December 2012.
- [22] A.V. Vieira and F.T. Toigo. Classificação BI-RADSTM: categorização de 4.968 mamografias. *Radiol Bras*, 35(4):205–208, 2002.
- [23] MAMAinfo. BI-RADS. BI-RADS in MAMAinfo, available in <http://www.mamainfo.org.br/texto.asp?c=bi-rads>, last access in 6 December 2012.
- [24] A. Canelas. BI-RADS, September 2009. Clínica Universitária de Imagiologia, Hospitais da Universidade de Coimbra, available in http://www.huc.min-saude.pt/imagiologia/biblio_data/BI_RADS.pdf, last access in 6 December 2012.
- [25] Banco de Saúde Lima, J.R. Mamografia - o que É BI-RADS ?, January 2011. Blog from Dr. J.R.Lima, available in <http://www.bancodesaude.com.br/user/3990/blog/mamografia-que-bi-rads>, last access in 6 December 2012.
- [26] W. Chiracharit, Yajie Sun, P. Kumhom, K. Chamnongthai, Charles. F. Babbs, and Edward J. Delp. Normal mammogram detection based on local probability difference transforms and support vector machines. *Transactions on Information and Systems (IEICE)*, E90-D(1):258–270, 2007. URL: http://search.ieice.org/bin/summary.php?id=e90-d_1_258.
- [27] M.Y. Elshinawy, A.A. Badawy, W.W. Abdelmageed, and M.F. Chouikha. Pre-CAD system for normal mammogram detection using local binary pattern features. In *IEEE 23rd*

- International Symposium on Computer-Based Medical Systems (CBMS)*, pages 352–357, 2010. URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6042669, doi:10.1109/CBMS.2010.6042669.
- [28] A.M. Sabu, N. Ponraj, and Poongodi. Textural features based breast cancer detection: A survey. *Journal of Emerging Trends in Computing and Information Sciences*, 3(9):1329–1334, 2012.
- [29] J. Kovacevic and M. Vetterli. Nonseparable multidimensional perfect reconstruction filter banks and wavelet bases for $r \times n$. *Information Theory, IEEE Transactions on*, 38(2):533–555, 1992.
- [30] R. Andrews and D.T. Nguyen. Separable versus quincunx wavelet transforms for image compression. In *6th IEEE int. Workshop on Signal Processing and Communication Systems, Melbourne, Australia*. Citeseer, 1998.
- [31] R.L. White. Methods for classification. *Methods for Classifications*, available in <http://sundog.stsci.edu/rick/SCMA/node1.html>, last access in 12 January 2013.
- [32] A. Statnikov, C.F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, 21(5):631–643, 2005.
- [33] K. Sai Deepak, N.V. Kartheek Medathati, and Jayanthi Sivaswamy. Detection and discrimination of disease-related abnormalities based on learning normal cases. *Pattern Recognition*, 2012. URL: <http://www.sciencedirect.com/science/article/pii/S0031320312001537>, doi:10.1016/j.patcog.2012.03.020.
- [34] S. Liu. *The analysis of digital mammograms: Spiculated tumor detection and normal mammogram characterization*. PhD, Purdue University, 1999. URL: <http://docs.lib.purdue.edu/dissertations/AAI9951984/>.
- [35] Y. Sun, C.F. Babbs, and E.J. Delp. Normal mammogram classification based on regional analysis. In *45th Midwest Symposium on Circuits and Systems (MWSCAS)*, volume 2, pages 375–378, 2002. URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1186876.
- [36] Y. Sun, C.F. Babbs, and E.J. Delp. Full-field mammogram analysis based on the identification of normal regions. In *IEEE International Symposium on Biomedical Imaging: Nano to Macro*, volume 2, pages 1131–1134, 2004. URL: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=1398742&contentType=Conference+Publications&searchField%3DSearch_All%26queryText%3Dfull+field+mammogram+analysis+based+on+the+identification+of+normal+regions.
- [37] Y. Sun. *Normal mammogram analysis*. PhD, Purdue University, 2004. URL: <https://engineering.purdue.edu/~ace/thesis/sun/sun-thesis.pdf>.
- [38] P. Taylor, S. Hajnal, M-H Dilhuydy, and B. Barreau. Measuring image texture to separate "difficult" from "easy" mammograms. *British Journal of Radiology*, 67(797):456–463, 1994. URL: <http://bjr.birjournals.org/cgi/content/abstract/67/797/456>.

- [39] Adele Cutler Leo Breiman. Random forests. Random Forests, available in <http://www.stat.berkeley.edu/~breiman/RandomForests/>, last access in 06 February 2013.
- [40] B. Sahiner, Heang-Ping Chan, N. Petrick, Datong Wei, M.A. Helvie, D.D. Adler, and M.M. Goodsitt. Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images. *IEEE Transactions on Medical Imaging (MedImg)*, 15(5):598–610, 1996. URL: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=538937&contentType=Journals+%26+Magazines&searchField%3DSearch_All%26queryText%3DClassification+of+mass+and+normal+breast+tissue%3A+a+convolution+neural+network+classifier+with+spatial+domain+and+texture+images, doi:10.1109/42.538937.
- [41] B.L. Kalman, W.R. Reinus, S.C. Kwasny, A. Laine, and L. Kotner. Prescreening entire mammograms for masses with artificial neural networks: Preliminary results. *Academic Radiology*, 4(6):405–414, 1997.
- [42] Jason Weston and Chris Watkins. Multi-class support vector machines. Technical report, Citeseer, 1998.
- [43] Koby Crammer and Yoram Singer. On the learnability and design of output codes for multi-class problems. *Machine Learning*, 47(2-3):201–233, 2002.
- [44] W. Chiracharit, Y. Sun, P. Kumhom, K. Chamnongthai, C. Babbs, and E.J. Delp. Normal mammogram classification based on a support vector machine utilizing crossed distribution features. In *26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEMBS)*, volume 1, pages 1581–1584, 2004. URL: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=1403481&contentType=Conference+Publications&searchField%3DSearch_All%26queryText%3DNormal+mammogram+classification+based+on+a+support+vector+machine+utilizing+crossed+distribution+features, doi:10.1109/IEMBS.2004.1403481.
- [45] J.N. Wolfe. Breast patterns as an index of risk for developing breast cancer. *American Journal of Roentgenology*, 126(6):1130–1137, 1976.
- [46] M.Y. Elshinawy, A. Badawy, W. Abdelmageed, and M. Chouikha. Effect of breast density in selecting features for normal mammogram detection. In *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 141–147, 2011. URL: <http://ieeexplore.ieee.org/xpl/articleDetails.jsp?reload=true&arnumber=5872374&contentType=Conference+Publications>, doi:10.1109/ISBI.2011.5872374.
- [47] M.Y. Elshinawyz, A.-H.A. Badawyy, W.W. Abdelmageedyy, and M.F. Chouikhaz. Comparing one-class and two-class SVM classifiers for normal mammogram detection. In *IEEE 39th Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–7, 2010. URL: <http://ieeexplore.ieee.org/xpl/articleDetails.jsp?jsessionId=ZwRkPgfl00JhrLG1BV1CksnnTLVW4TSZ1MnygDtDcHxSJhzgf3xn!-2011277867?arnumber=5759708&contentType=Conference+Publications>, doi:10.1109/AIPR.2010.5759708.

- [48] M.Y. Elshinawy, A.H.A. Badawy, W.W. Abdelmageed, and M.F. Chouikha. Normal mammogram detection using density information and texture features. In *SiiM*, 2011. URL: http://www.siiim2011.org/abstracts/advanced_visualization_tools_ss_elshinawy.html.
- [49] I. Domingues, J.S. Cardoso, and P. Cardoso. Identification of benign breasts during mammogram screening. In *18th edition of the Portuguese Conference on Pattern Recognition (RecPad)*, 2012.
- [50] S. Liu, C.F. Babbs, and E.J. Delp. Normal mammogram analysis and recognition. In *International Conference on Image Processing (ICIP)*, volume 1, pages 727–731, 1998. URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=723599&tag=1.
- [51] F. Eddaoudi, F. Regragui, and K. Laraki. Characterization of the normal mammograms based on statistical features. In *Second International Symposium on Communications, Control and Signal Processing (ISCCSP)*, 2005. URL: <http://www.eurasip.org/Proceedings/Ext/ISCCSP2006/defevent/papers/cr1086.pdf>.
- [52] J.N. Wolfe. Risk for breast cancer development determined by mammographic parenchymal pattern. *Cancer*, 37(5):2486–2492, 1976.
- [53] N.F. Boyd, J.W. Byng, R.A. Jong, E.K. Fishell, L.E. Little, A.B. Miller, G.A. Lockwood, D.L. Tritchler, and M.J. Yaffe. Quantitative classification of mammographic densities and breast cancer risk: results from the canadian national breast screening study. *Journal of the National Cancer Institute*, 87(9):670–675, 1995.
- [54] K. Bovis and S. Singh. Classification of mammographic breast density using a combined classifier paradigm. In *Medical Image Understanding and Analysis (MIUA) Conference, Portsmouth*, 2002.
- [55] S. Petroudi and M. Brady. Breast density dependent computer aided detection. *Digital Mammography*, pages 34–38, 2006.
- [56] A. Oliver, J. Freixenet, R. Marti, J. Pont, E. Pérez, E.R.E. Denton, and R. Zwigelaar. A novel breast tissue density classification methodology. *Information Technology in Biomedicine, IEEE Transactions on*, 12(1):55–65, 2008.
- [57] I.C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M.J. Cardoso, and J.S. Cardoso. Inbreast: Towards a full field digital mammographic database. In *Academic Radiology*, volume 19, page 236–248, 2012.
- [58] Mathworks. `imadjust` function help, documentation center - mathworks. MatLab help to `imadjust`, available in <http://www.mathworks.com/help/images/ref/imadjust.html>, last access in 2 March 2013.
- [59] J.S. Cardoso, I. Domingues, I. Amaral, I. Moreira, P. Passarinho, J. Santa Comba, R. Correia, and M.J. Cardoso. Pectoral muscle detection in mammograms based on polar coordinates and the shortest path. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pages 4781–4784. IEEE, 2010.
- [60] Mathworks. `Graycomatrix` help, documentation center - mathworks. MatLab help to `GLCM`, available in <http://www.mathworks.com/help/images/ref/graycomatrix.html>, last access in 05 April 2013.

- [61] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59, 1996.
- [62] M. Heikkilä and T. Ahonen. Department of computer science and engineering, university of oulu. Matlab Code, LBP, available in <http://www.cse.oulu.fi/CMV/Downloads/LBP Matlab>, last access in 05 April 2013.
- [63] Matti Pietikäinen. *Computer vision using local binary patterns*, volume 40. Springer, 2011.
- [64] Cornell University. Instance-based learning. Instance-Based Learning Course, available in http://www.cs.cornell.edu/courses/cs472/2005fa/lectures/7-knn_6up.pdf, last access in 04 June 2013.
- [65] Cunningham, Pdraig, Delany, and S. Jane. k-nearest neighbour classifiers. *Multiple Classifier Systems*, pages 1–17, 2007.
- [66] Mathworks. svmtrain function help, documentation center - mathworks. MatLab help to svmtrain, available in <http://www.mathworks.com/help/stats/svmtrain.html>, last access in 04 June 2013.
- [67] Mathworks. Svmclassify function help, documentation center - mathworks. MatLab help to svmclassify, available in <http://www.mathworks.com/help/stats/svmclassify.html>, last access in 04 June 2013.
- [68] London Tae-Kyun Kim, Imperial College. Random forest and committee machine - lecture 7 - 8. Slides Lecture from the Imperial College, London, available in http://www.iis.ee.ic.ac.uk/icvl/mlcv/lec78_pub.pdf, last access in 11 Jun 2013.
- [69] Togaware Graham Williams. Random forests. Random Forests explained by Graham Williams, available in http://datamining.togaware.com/survivor/Random_Forests.html, last access in 11 Jun 2013.
- [70] Mathworks. treebagger function help, documentation center - mathworks. MatLab help to treebagger, available in <http://www.mathworks.com/help/stats/treebagger.html>, last access in 25 May 2013.
- [71] Mathworks. predict function help, documentation center - mathworks. MatLab help to predict function, available in <http://www.mathworks.com/help/stats/treebagger.predict.html>, last access in 25 May 2013.