U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

MSc

2.º CICLO

FCUP
2014

U.PORTO

Multiple Kernel Learning for Breast Cancer Classification

Gustavo Barbosa Augusto

FC

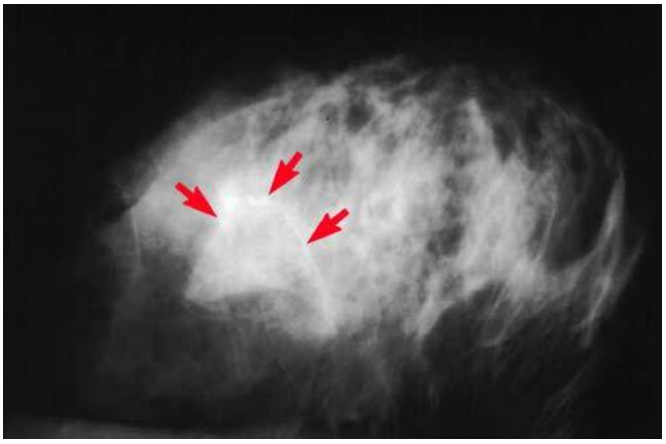
Multiple Kernel Learning for Breast Cancer Classification

Gustavo Barbosa Augusto

Dissertação de Mestrado apresentada à
Faculdade de Ciências da Universidade do Porto em
Ciência de Computadores

2014

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO



Multiple Kernel Learning for Breast Cancer Classification

Gustavo Barbosa Augusto
Mestrado em Ciência de Computadores
Departamento de Ciência de Computadores
2014

Orientador
Inês Dutra, Professora Doutora, DCC, UP

Coorientador
Ricardo Sousa, Doutor, Instituto de Telecomunicações

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____

Acknowledgements

First, I want to give a special thanks to my supervisor Inês Dutra and to my co-supervisor Ricardo Sousa for their support, for giving me constant feedback and for pushing me forward all the time, it really kept me motivated during this work! I also want to thanks my family and all my dear friends who helped me to relax during this stressful time.

Resumo

O cancro de mama é uma das maiores causas de morte de cancro em mulheres de acordo com a organização mundial de saúde. Logo, medidas preventivas são necessárias para reduzir a percentagem de morte, medidas tais como o rastreio da mamografia. Este exame permite a detecção do cancro nos seus estados iniciais. Entretanto, a análise de mamografias tem um custo elevado, devido à necessidade de um radiologista ter que detectar e classificar manualmente anomalias nas imagens. Como resultado, sistemas informáticos de apoio ao diagnóstico do cancro de mama têm sido utilizados com o intuito de reduzir o custo das análises das mamografias e para também aumentar o sucesso na classificação das anomalias, porque até os profissionais cometem erros na classificação de anomalias. Neste trabalho, analisamos o estado actual da literatura, apresentando neste sentido quais: as bases de dados com imagens de mamografias estão disponíveis, os métodos mais comuns utilizados no processamento de imagem no cancro de mama e seus respectivos métodos de classificação. Para o reconhecimento de patologias, foi utilizado o Multiple Kernel Learning (MKL) que tem demonstrado superioridade em relação ao Support Vector Machine (SVM) e no contexto do cancro de mama, poderá resultar numa melhor qualidade nos diagnósticos. Para provar que essa superioridade também existe com dados de imagens médicas no contexto do cancro de mama, fizemos um estudo comparativo entre o SVM e o MKL usando dados obtidos através dos métodos de processamento de imagem mais populares na literatura. Concluímos assim que o método MKL ultrapassa o estado da arte em classificação de cancro de mama usando apenas casos de massas sem a utilização de dados clínicos, obtendo uma Area Under the Receiver Operating Curve (AUC) de 0.871 e também em classificação de cancro de mama usando casos com todos os tipos de anormalias e também utilizando dados clínicos obtendo uma AUC de 0.834.

Abstract

Breast cancer is the most common cause of cancer death among women according to the World Health Organization. Thus, preventive measures are required to reduce death rate, which include for instance, a screening mammography of the patient. This allows the detection of the cancer in early stages. However, such analysis is expensive, as it requires a radiologist to detect and classify anomalies in the breast image. As a result, computer aided diagnosis systems of breast cancer classification have been used to reduce the cost of the mammogram analysis and to increase the success ratio of the classification since even professionals make mistakes on anomaly classification. In this work, we analyse the current extensive literature on this field, thus reporting currently available breast image databases and most commonly used image processing methods and the respective classification methods. For the pathologies detection, we used the Multiple Kernel Learning (MKL) which has demonstrated superiority in relation wiht the Support Vector machine (SVM) and in the context of breast cancer, it could also result in a better quality of diagnosis. In order to prove if the MKL remains superior to SVM using breast cancer image data, we perform a comparison study between SVM and MKL using features from the most popular image processing methods on the literature. Based on this study, we conclude that our method surpasses the state of the art on breast cancer mass classification without clinical data with an Area Under the Receiver Operating Curve (AUC) of 0.871 and on breast cancer classification for all findings with clinical data with an AUC of 0.834.

Palavras Chave / Keywords

Palavras Chave:

- Diagnóstico supervisionado
- Cancro de mama
- Aprendizagem Automtica
- Processamento de imagem

Keywords:

- Computer Aided Diagnosis
- Breast Cancer
- Machine Learning
- Image Processing

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Breast cancer is the most common cause of cancer death among women according to the World Health Organization [5]. One-third of breast cancer deaths could be avoided if the breast cancer was detected and treated on early stages. The process of detection and classification of breast cancer is based on images obtained by Magnetic Resonance Imaging (MRI), mammography (x-ray images of the breast) and Ultra sound images where a specialist analyses the breast imaging to detect any anomaly and classify it as malign or benign [6]. The x-ray is the source for breast imaging available where radiologists are trained to detect and classify breast cancer. However, breast cancer detection and classification by a radiologist using mammograms is not a flawless technique. Some support measures are taken when possible, such as a second reading of the mammogram by another radiologist or the usage of Computer Aided Diagnosis (CAD) systems. It has been proven that CAD systems can outperform a second reading [7]. Most CAD systems are based on two types of information, the mammogram image and background knowledge of the patient, i.e. clinical data which will be described in Chapter 3. Despite their effectiveness, increasing CAD breast cancer classification ratio is still an active research topic. Many methods have been applied to breast cancer classification such as CAD systems using machine learning methods like K-Nearest neighbours (kNN), Support vector machines (SVMs) explained in Chapter 2. One promising approach is Multi Kernel Learning (MKL) [8], which uses the SVM method to each feature being able to analyse them separately. To the best of our knowledge, MKL has not yet been applied in the domain of breast cancer classification. In this work, we investigate the behaviour of MKL applied to image features and clinical annotated data in comparison with other learning methods for breast cancer classification. To perform this study we test different learning methods

and image features. After the study we could be able to create a system using MKL to classify breast cancer images between benign and malign. This system could be divided in two main components: The first component is the image processing part in which digital information from breast images are captured, such as intensity changes, shapes and size with a breast lesion marked by a radiologist. The second uses the processed information to create the MKL. Resulting in the introduction of MKL for breast cancer classification with the purpose of increasing the overall performance. In this chapter we discuss the main challenges of breast cancer detection by image processing, how a breast cancer system can be created, how it works and why it should be used.

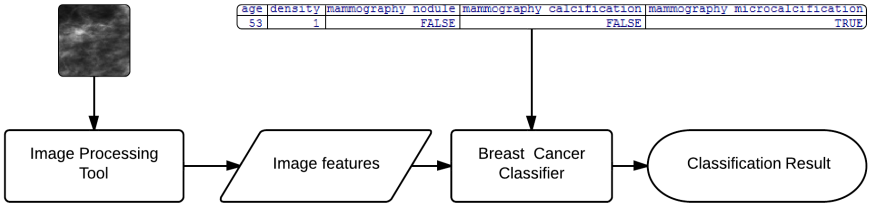


Figure 1.1: Represents the system overview. The system receives as input a mammogram image cropped by region of interest (region that contains the finding), and the respective clinical data. Then, it retrieves from the image features calculated by the image processing module. Image features and clinical data are merged to the breast cancer classifier to predict if the finding is benign or malign.

1.1 Motivation

Encouraging results have been reported in the literature reinforcing the potential of CAD systems [9, 10, 1]. Breast cancer classification is an important task that needs to be improved in order to spare patients of unnecessary procedures and to be able to properly treat patients with malignant cases. We had access to two breast cancer repositories. The availability of these datasets and the proximity with breast cancer experts added an extra motivation to the realization of this work. In this thesis we are addressing two major challenges regarding CAD for Breast cancer detection:

- **Discriminative image features** is essential for breast cancer classification. Every year new approaches are designed in order to extract breast cancer features from a mammography. Since those approaches are not perfect, it is still an ongoing challenge to find the best set of features that fully describe a cancer finding in a mammogram.

- **Improve breast cancer classification performance** is a hard challenge, although image features play a big role in the classifier performance, studies are necessary in order to create a model that can learn as much as possible from the feature set. The average performance of breast cancer CAD systems in the literature of 85% [1] leaving space for improvements.

One of the most used classifiers for malignancy breast cancer classification is the SVM, due to its simplicity and overall good performance. Some other more elaborated methods such as MKL are not usual due their non trivial usage. Given the fact that the MKL has been proven to surpass the SVM in some cases[8], we are motivated to explore in this thesis the advantages of MKL as an approach for breast cancer classification aiming to report improved performance for breast cancer classification in comparison with the SVM method. Even though the purpose of this thesis is to explore the learning capability of the state-of-the-art machine learning methods for breast cancer learning, the benefits of this experimental study, is also extensible to the research community, where researchers can replicate the feature extraction or classification methods used or us the obtained results for their comparison experiments.

1.2 Objectives

This thesis presents an experimental study to evaluate the performance of MKL using images containing clinical data and image descriptors from ROIs containing lesion of a mammogram as input to return the binary result malign or benign. The experiment is illustrated in Section 1.1. In order to accomplish the objectives described in the subsection below, we designed a model for the experiment that was divided in three modules as we can see in Section 1.2. First, the database module which will contain the datasets that the user want to study with the respective clinical data and pre-generated features from other modules. Second the image processing module which allows to generate image features depending on the image processing methods added. At last there is the classification model, which allows to create several classifiers and perform the same performance experiments to obtain the best parametrization and compare their efficiency. Both the MKL and the experimental study with MKL are described in more detail in Chapter 4 and Chapter 5.

The main goal of this thesis is to create a system classify breast cancer lesions with high performance, therefore the following objectives were proposed:

- **Research** of the current state of the art of computer aided diagnosis for breast cancer;

- **Test** image descriptors that retrieves high quality features from breast cancer images regarding the lesion class;

- **Use** an library containing a MKL classifier that will retrieve as much information as possible for each type of features;

- **Deploy** a system to classify images of breast cancer leisures between benign and malign.

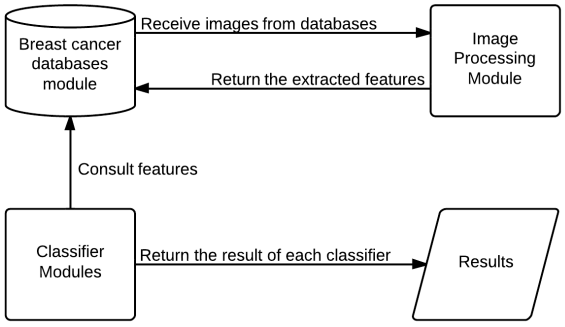


Figure 1.2: Represent the experiment model overview, where each module communicate as follows: The Database module provides breast cancer images to the Image processing modules, and the image processing module return the respective generated features. The Database also keep the generated features. The classifier module request the image generated features and clinical data to perform an experiment that will test several classifiers

The main features of the proposed experiment are:

- **Person Independent** - Able to classify breast lesions of all 4 types of breast density;

- **Lesion Independent** - Able to classify different breast lesions such as calcifications and masses;

- **Multiple Image Descriptors Extracted** - Able to analyse several image descriptors extracted from the breast cancer mammogram datasets.

1.3 Applications

In the course of our work we adapted the system to create two applications to study and classify breast patches which is the objective of this thesis. They are:

1. **Database Analyser:** This tool can receive an set of breast cancer images and an outline and return a text file containing the data generated by the system image descriptors.
2. **Breast Cancer Classifier:** This application receives an image descriptor from an ROI of a breast cancer image as input and shows the classification of the lesion if its benign or malign.

1.4 Contributions

This section describes the contributions that were obtained during this research. These include the studies and system developed in this thesis process.

- A review of the state of the art of breast cancer classification.
- We present new results using MKL as a robust alternative for breast cancer recognition.

1.5 Outline

The remaining chapters of this thesis are organised as follows:

Chapter 2 Describes the background knowledge behind breast cancer systems;

Chapter 3 Presents an overview of the studies and the current methodologies for breast cancer CAD systems;

Chapter 4 Includes a thoroughly description of the MKL;

Chapter 5 Describe the experiment study, discuss the results obtained with the CAD system developed and analyse its strengths and flaws in comparison with recent systems;

Chapter 6 Discusses the presented work and the future directions.

Chapter 2

Concepts and Definitions

CAD for breast cancer systems are built using multi disciplinary knowledge. In this chapter we expose some base concepts in diverse areas such as computer vision, machine learning and radiology, for self-contained purposes of this thesis before we start analysing current work in the literature.

2.1 Computer Vision

Computer vision is the area of research whose goal is to devise algorithms to process and understand images/videos. Humans are proficient at understanding what they see. For instance, if they look at a room and asked to search for a table they can clearly identify its location and the colour of the material. Despite being an effortless task for humans, for computers it may be a really hard task. Why is it so hard? The computer must have all the information necessary to conclude what is a table. That information should contain, what kind of shapes a table can have, which kind of texture they can have, which sizes they can have and so on. (This process encompasses on highly computational intensive processes since image capture until recognition) The following paragraph will explain in more detail how those two parts are performed and which methods are used.

There are several ways to process an image, in general the chosen processing method depends on the objective of the application. A colour image is represented in a 3 dimensional matrix containing colour values between 0 to 255 (in the case of 8 bits) for each pixel coordinate. The third coordinate corresponds to the colour information,

two examples of colour representation are: Red, Green, Blue (RGB) or Intensity (Greyscale). Depending on the size of the image and depending on the domain of image (e.g RGB, Grayscale) there may be too much information to work with, i.e. an image would result in a feature vector of size n, where n = width * height * 1 (in case of the Grayscale) which gives only the spacial and intensity information of each pixel since it contains only pixel values and their respective coordinates. In order to obtain better feature vectors with more information and in some cases to reduce this feature vector, descriptors are created.

There are several kinds of descriptors depending on the type of image that will be processed. Some of them use only the intensity values such as statistical descriptors containing histograms of intensity, others change from the intensity domain to the frequency domain in order to efficiently perform filters to enhance the image for a certain purpose and also to be able to split low frequencies (image texture and shape) and high frequencies (image edges and detail), there are also bandpass filters (filter a range of frequency) such as Gabor wavelets used as a descriptor and as base in several computer vision methods. For pattern recognition and matching there are the local invariant descriptors that allows the computer to find in interesting points and extract a feature vector that describes those points or regions, these methods use the location of the points in the matrix, their neighbourhood. In more detail those methods are described below.

1. **Statistical Intensity Patterns:** There are two main types of statistical information that may be extracted from an entire image or from a component of a image (e.g. if we want to process an component inside the image, we may want to first crop this region of interest (ROI) containing the component and analyse only the information inside that ROI). A first type is the single pixel approaches based on the histogram of intensity of an image i.e. a vector containing the frequency of each intensity from 0 to 255, or based on the mean (average of intensity of all pixels), the standard deviation, maximum and minimum intensity value and the skewness. Gray level co-occurrence matrix (GLCM) [11] they are based on the neighbourhood of each pixel, by calculating metrics such as homogeneity, energy, average, entropy, smoothness and correlation. The latter can capture information of higher order from the image. Works have demonstrated the robustness of this method for texture recognition, one example may be breast cancer recognition [12, 1]. However, the intensity domain enlighten only few part of the information an image may contain.

2. **Frequency Domain:** In signal processing, it was developed the transformation of Fourier and their inverse which can be used in image processing to swap between frequency or intensity domain (in this case also known as spatial domain). The interesting part of the frequency domain is that high frequencies contain information regarding edges and detail, for example, high pass Gaussian and homomorphic filter. On other hand if the goal is to detect shapes and texture, removing high frequencies might help, for this purpose there are low pass filters like Gaussian blur. Detailed information regarding the Fourier transform, frequency domain and filters can be found at [13]. We next explain in more detail the Gabor Wavelet Descriptor.
   The Gabor function was proposed at 1946, and its frequently used in image feature description. According to [14], in a one dimensional case, a Gabor function can be defined as a complex exponential localized around x = 0 by the envelope of a Gaussian window shape represented by Eq. (2.1) for each $\alpha \in \mathbb{R}^+$ and each ξ,x $\in \mathbb{R}$, where $\alpha = (2\sigma^2)^{-1}$, σ^2 is a variance and ξ is a frequency. In two dimensional cases, the function is separable into a series of one dimensional functions. The elements of a family of mutually similar Gabor functions are called wavelets when they are created by dilation and shift from one elementary Gabor function, the mother wavelet that can be defined in Eq. (2.2), for a $\in \mathbb{R}^+$ (scale) and b $\in \mathbb{R}$ (shift). In [14] its also demonstrated how they can be used for blob detection (detection of regions formed by points that are similar to each other), corner detection (regions that form a corner) that are used in some Local Invariant Descriptors. For the interested reader, more information regarding descriptors in the frequency domain can be referred to [13].

$$g_{\alpha,\xi}(x) = \sqrt{\alpha/\pi} e^{-\alpha x^2} e^{-i\xi x} \tag{2.1}$$

$$g_{\alpha,\xi,a,b}(x) = |a|^{-1/2} g_{\alpha,\xi}(\frac{x-b}{a}) \tag{2.2}$$

3. **Spacial Image Analysis:** Spacial image analysis consists in two components: Identification of interest points and description based on local information. Interesting points may be corners detected by the Harris Corner Detector or blobs which can be found using a Hessian matrix. A detailed comparative overview of the local features explaining each method can be found at [15]. An ideal interest point should be invariant to scaling, orientation, affine distortions and illumination changes (i.e. an interest point should match even if any of those properties change e.g. in the case of the scaling property, if take two pictures of a monument, one with zoom and other without, the same interest point in

both images should match), although the necessity of each characteristic may vary according to each application scenario, e.g. if in our application scenario its ensured that all pictures are taken on the same place with the same distance, the method does not require to be scale invariant. One of the most popular local invariant descriptor is the Scale-Invariant Feature Transform (SIFT)[16]. This method is invariant to uniform scaling, orientation, rotation and partially invariant to affine distortions and illumination changes. In more detail the SIFT descriptor is illustrated at Figure 2.1 where we want to detect the scale-space extrema to obtain scale invariance using Difference of Gaussian[16] to find the local maxima across the scale and space which gives us a list of (x,y,σ) values containing a potential keypoint at (x,y) at σ scale. To avoid capturing noise, an heuristic approach to select only points of interest. Then, the orientation of each keypoint is calculated, i.e. the gradient magnitude and direction is calculated. Then, the keypoint final descriptor is obtained from a 4×4 block estimated from 8 orientations. Resulting in a total of 128 bin values.

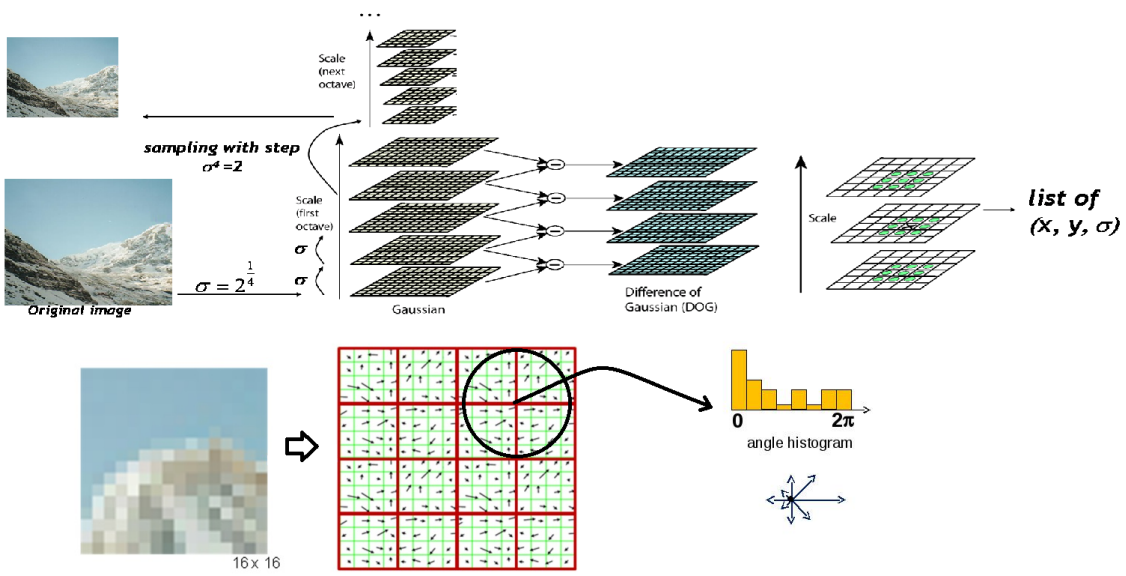


Figure 2.1: The SIFT feature vector extraction. On top it is extracted the σ for each keypoint. On the lower part for each interest keypoint the gradients are calculated to obtain the orientation matrix and the angle histograms and the keypoint descriptors are extracted for each quarter.

2.2 Machine Learning

Machine learning is a field of research whose goal is the design of automatic learning systems. But why is it useful in computer vision? As mentioned previously, we can extract so much information from an image, but with that information alone it is really hard for human interpretation. Therefore, there is a necessity for automatic methods that can automatically uncover patterns in that data and interpret them in order to create conclusions and even predict conclusions regarding unknown new data. In general, data can be represented in qualitative and quantitative measures, such as integer values and real numbers, or ordinal variables (e.g. quantities "few", "several" and "lots") and nominal variables (e.g. round, square, triangle). The storage of a set of vectors of data (also known as feature vector) is called a "dataset".
There are two main types of learning. First the supervised learning type where we have a dataset containing feature vectors where each vector is mapped to a label that describes the vector (e.g. a vector with t-shirt measurements such as length and width mapped with size labels such as small, medium and large.). This dataset is called "train set", with it we create models that may predict new labels for new feature vectors. The problem of prediction may vary according to the label type of data, for instance if the label is nominal, it is called a "classification" problem, if the label is an ordinal its called a "regression" problem and if the label is a ordinal type its called "ordinal regression". Second the unsupervised type where given some data the goal is to find interesting patterns, these types of problems are usually less accurate and unknown regarding error metrics since there is no comparative label.
A **classification** problem can be formalized as a function f that receives a feature vector v and returns the v respective class. If the number of the classes equals two, its called a binary classification, otherwise it is a multiclass classification problem. Several approaches were taken to create a classification model, like distance formulas such as kNN [17], decision trees [18] and inductive logic programming, naive Bayes method [19], and other different approaches like support vector machines [20]. Most of these methods are popular and widely used, we explain them in some more detail below.

- **K-Nearest Neighbours [17]:** This is the simplest decision method. It considers the whole dataset as the training model. The prediction of a row is obtained by returning the most frequent class of the K nearest rows given a distance metric. In most of the cases the distance metric used is the Euclidean Distance. The Euclidean distance of two feature vectors R1 and R2 can be represented as

in Eq. (2.3) where n is the number of features of each vector. The returned output corresponds to the most frequent output provided by the k rows.

$$Distance(R1, R2) = \sqrt{(R1_1 - R2_1)^2 + (R1_2 - R2_2)^2 + \cdots + (R1_n - R2_n)^2} \tag{2.3}$$

- **Classification and Regression Trees (CART):** This supervised method aims to grow a tree by creating branches with the purpose of splitting one class from others until there is not enough information to separate the dataset or a stopping criterion is reached. This tree is composed by three elements, first it contains a root node which has no parent node. Second, a tree may have internal nodes. These nodes have exactly one parent node and have two child nodes, and at last the tree may have leaf nodes. These nodes have one parent node and zero child nodes.
  The tree works as it follows: Given a feature vector from a dataset and starting from the root node, a logical test is performed. We move to the child node that agrees with the logical test result and repeat the process until the child node is a leaf node. The leaf node contains the predicted class for the given feature vector. This logical test uses one feature from feature vector of the train set which can be numerical (e.g. Size <40) or nominal (e.g. colour == red) depending on the type of the feature. Logical tests are created, i.e. new branches on the tree are formed, until the tree contains enough paths of logical tests that properly classify as many rows as possible from the train set.
  An example of a tree model is illustrated in Figure 2.2. A fictional cancer classification tree based on the patient information, the model itself is visually understandable, however the more complex the model is the harder will be for humans to visually understand the whole model. A tree model can also be used for a regression problem, instead of separating by classes it would separate by regression values (numerical).
  As we saw previously, a classification tree is formed by logical tests and a good test would be a test that can create a pure node, i.e. a test that can split all cases of a certain class from the cases of other classes. One way to evaluate how many tests a tree should have and how to select the best tests to construct a tree, would be to create logical tests and see if they increase or decrease the error rate of the tree by keeping the tests that decrease the error rate and repeating this empirical process until a desired error rate is achieved. The error rate is calculated by 1 - accuracy (explained in Eq. (2.8)), there are also two other measures used to evaluate the overall purity of the nodes such as the Gini Index

and the Entropy. They can be respectively represented by the Eq. (2.4) and Eq. (2.5) where $p(\mathrm{i}|\mathrm{t})$ is the fraction of recordings belonging to class i at a given node t (for examples and full algorithms regarding growing decision trees and evaluating node purity read [21]. However, finding the optimal decision tree for data was also classified as a NP-Complete problem [22]. Therefore, greedy methods were used to select the best parameters and tests to grow good trees instead of searching for the optimal tree, one of the famous methods used in the most popular data mining tools (such as R and Weka), is the CART algorithm proposed by [18]. After growing a tree, there may be too many branches, causing a problem known as "overfitting" of the data, i.e. too many specific rules for the training data that increase the overall training accuracy but tend to decrease the accuracy of the classifier when using data outside the training set. In order to solve this problem pruning methods are used after growing a tree in order to reduce branches that do not increase the impurity of other nodes by a certain threshold.

$$Gini(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2 \tag{2.4}$$

$$Entropy(t) = -\sum_{i=0}^{c-1} p(i|t) \log_2 p(i|t) \tag{2.5}$$

There are also several methods based in decision trees called ensemble methods, they use the concept of bagging [23], by creating several weak decision tree classifiers and use them to cover each other weaknesses resulting in a stronger classifier. One famous ensemble method is the Random Forest [24], it is widely used for the classification problem obtaining great overall performance comparable to one of he most used classifiers, the SVM.

- **Support Vector Machine (SVM):** This method was introduced in 1995 by Vapnik and Cortes [25]. Nowadays, it is widely used to solve most of the classification and regression problems due to its good overall performance dealing with any data. The basic concept of the SVM is to find an hyperplane that splits one class from other classes within the data feature space, as we can see in Figure 2.3. There may be an infinity of hyperplanes for this purpose. In order to select the best hyperplane, first we find the margins hyperplanes (m1 and m3 in 2.3) for a certain class and for the rest of the data, then we calculate the maximum margin hyperplane (m2). This method will be analysed in depth on Chapter 4.

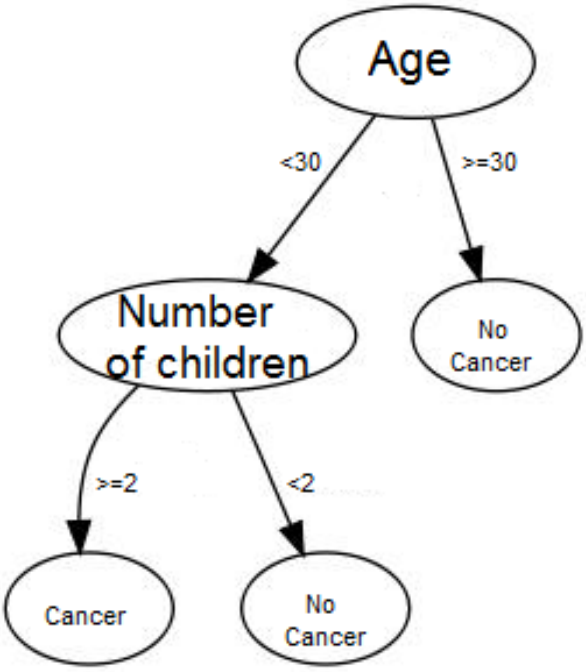


Figure 2.2: A classification tree created to classify between cancer or no cancer given the patient information.

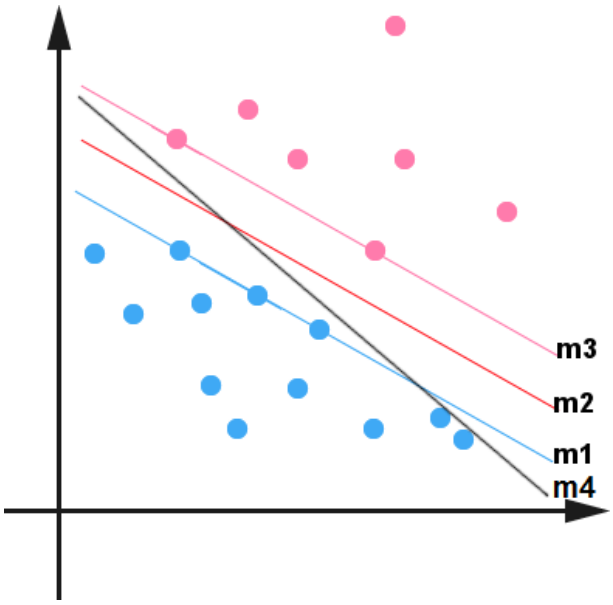


Figure 2.3: Four hyperplanes that divides a bi-dimensional feature space. Where m2 is the maximum margin hyperplane of this feature space because its in the middle of m1 (formed by 3 support vectors) and m3 (formed by 2 support vectors), which are the hyperplanes with minimum margin of each class. The m4 is an example of a non optimal possible hyperplane that divides the data from each class.

- **Naive Bayes:** Bayesian classifiers are statistical classifiers, they predict the probability of a data row belonging to a class. A particular class of Bayesian classifiers which often obtain good results [26] is the Naive Bayes classifier. Based on the Bayes Theorem Eq. (2.6), where D is a dataset that contains feature vectors x with size t, where: x is labeled with a class C_i; $_i$ vary according to the number of classes from D; H_i is the hypothesis that states that a certain test case x belongs to a class from C; P(H) is the prior probability of the hypothesis H; P(x|H) means the likelihood, i.e. the conditional probability of x happening knowing P(H); P(H|x) is the posterior probability of H knowing x; P(x) is a normalization constant that does not affect the decision.
  The Naive Bayes predict the class of a new case by returning the class with the highest P(x|H_i).P(H_i). Since the correct computation of P(X|H_i) would be complex, the Naive Bayes simplify it by "naively" assuming that all hypothesis H_i for each class are independent. Allowing to calculate the P(x|H_i) in Eq. (2.7). However, in some cases P(x_k |H_i) of a certain k can be zero which would affect all other cases x of this H_I, in order to overcome this issue there is an additive smoothing method which adapt the model in order to avoid null probabilities. For additional information consult [19].

$$P(H|_x) = \frac{P(_x|H)P(H)}{P(_x)} \tag{2.6}$$

$$P(x|H_i) = \prod_{k=1}^{t} P(x_k|H_i) \tag{2.7}$$

2.3 Performance Assessment

These machine learning methods are very popular, but how can they be evaluated? In this section we will present a set of common measures that will be used in this thesis. Given a learning model to predict a set of instances, we record each predicted class and we compare with the true class. Model performance assessments can be decomposed in: True positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) as represented in Table 2.1, this matrix is called confusion matrix. Several performance metrics are based in this matrix values.

Table 2.1: Confusion matrix

		Ground truth values	
		Positive	Negative
Predicted values	Positive	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)

- **Accuracy:** measures the percentage of the correct predicted values, represented in Eq. (2.8). The opposite of accuracy (1 - accuracy) is known as **error rate**.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{2.8}$$

- **Recall:** also known as **sensitivity**, measures the number of positive cases that were properly predicted as positive Eq. (2.9).

$$Recall = \frac{TP}{(TP + FN)} \tag{2.9}$$

- **Precision:** measures the proportion of predicted positive cases that were correct Eq. (2.10).

$$Precision = \frac{TP}{(TP + FP)} \tag{2.10}$$

- **Specificity:** measures the number of false cases that were properly predicted as false. Eq. (2.11).

$$Specificity = \frac{TN}{(TN + FP)} \tag{2.11}$$

- **F-Measure:** is the harmonic mean between precision and recall useful for ranking or comparing methods Eq. (2.12).

$$F - Measure = 2 \times \frac{(precision * recall)}{(precision + recall)} \tag{2.12}$$

- **Receiver Operator Characteristic (ROC) Curve:** is the visual representation of the sensitivity and specificity distributions obtained by changing the decision threshold of a model [27], illustrated in Figure 2.4.

- **Area Under Curve ROC (AUC or AUCROC):** measures the area of a ROC curve to evaluate the classification performance. The area value can be

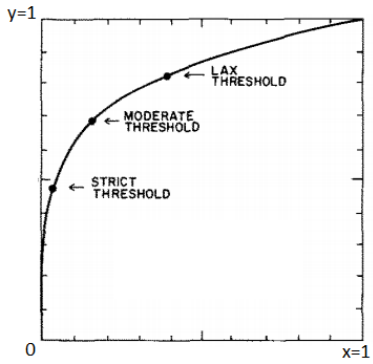


Figure 2.4: A ROC curve, where the x = 100 - specificity and the y = sensitivity

interpreted as the probability of a randomly selected positive sample will rank higher than a randomly selected negative sample [28].

However these metrics alone cannot ensure their statistic rightness, in order to achieve statistical validation the following methods were created.

- **Holdout:** This evaluation method consists in separating the dataset in two different subsets using random subsampling without repetition, i.e. selecting entries from the dataset randomly without repeating selected entries. For instance, we can use 70% of the original dataset to construct a learning model (train dataset) and the remaining 30% are kept for testing the model (test dataset).

- **K-fold cross validation:** This method is most used for classifier parameter optimizations. The k-fold cross validation goal is to divide the dataset em k subsets, and use each k subset as test set and the remaining k-1 datasets as train dataset. After running all k subsets, as final result its calculated the average of the evaluation metrics calculated for each k subset. According to some studies [29], the value k = 10 holds better sensitivity for cross validation.

In order to increase statistical significance, these methods are usually repeated several times (e.g. 50 to 200 times, an arbitrary number where the results and the standard deviation converge) and the average of the repetitions is calculated as final result in other to avoid biased results, i.e. results that are induced by the overfitting of the data. The results may vary on each repetition according to the partitioning of the dataset by the selected validation method. To reduce this variance, some stratification measures may be taken, i.e. ensure the trainset and testset have the same ratio of entries for each class. For deeper understanding of how to calculate the evaluation

metrics within these evaluation methods check the study created by [30], they analyse the performance, the bias and variance of these metrics within cross validation and explain how they should be implemented in order to properly compare the results obtained in an experiment.

2.4 Application Domain

In medicine, early breast cancer diagnosis by radiologists using mammograms (visual analysis of x-ray images obtained during mammography) is a hard task. Often requiring more than 1 professional to increase the diagnosis accuracy. However, having more than 1 professional doubles the cost of the analysis of the mammogram. And even with second opinions there are still false positives diagnosis where unnecessary diagnosis and treatments are performed, and there are also false negatives which results in giving the patient a false feeling of security, where the patient goes home without proper treatment. With the objective to reduce the cost of breast cancer diagnosis and to reduce the misdiagnosis, overdiagnosis and overtreatment issues, many researchers study the application of **Computer vision systems** to assist in the diagnosis of a patient, where the system is used as a second opinion to confirm the diagnosis rightness. In order to understand how computer vision can be applied in breast cancer classification, we should learn which are the terms and procedures in mammography screening. To know what kind of input the computer vision system will receive, what kind of problems they are trying to find with the given input and which result is comprehensible in the area of breast cancer mammogram diagnosis.

2.4.1 Screening mammography

Screening mammography is a medical exam to search for breast cancers in asymptomatic patients, mainly recommended annually or biennially to women with 40 years old or more, by the US National Cancer Institute (among other cancer institutes worldwide). This exam typically requires to position each breast in two different positions in order to take two views of each breast. One view from above which is called cranial-caudal (CC) and the second view from an oblique or angled view called mediolateral-oblique (MLO), these views provide visualization of the breast tissue in 2 planes for cancer detection. There are also two types of mammographies, the conventional screen-film mammograms and the full-field digital mammography

(FFDM)[31]. According to [32] digital mammography is significantly better than conventional mammography in detecting cancer in young women, premenopausal and perimenopausal women, and women with dense breasts. However, and due to screening costs, some hospitals may not have their system updated with the FFDM, despite their improvement mammogram diagnosis. The breast positioning in both CC and MLO views and their respective screen-film mammogram and digital mammogram are illustrated in Figure 2.5. More information regarding positioning of the patient and the mammography procedure can be found at [33]. After processing the mammograms, a diagnosis mammography is applied when anomalies are found or when the patient have symptoms regarding the breast [34]. In the following subsections we will explain the anomalies that can be found at a mammogram and how they can be classified.

2.4.2 Breast anomalies

In a mammogram, several anomalies can be found, however not all of them are linked to cancer or neither represent a threat. Between those anomalies there are masses, calcifications, architectural distortion, asymmetries, intramammary lymph nodes, skin lesions and dilated ducts, which are described bellow.

- **Masses:** are formed by breast tissue or they can also be cysts formed by benign collections of fluid in the breast. They can be detected by mammography in most cases years before they get large enough to be detected by touch. In this finding its important to take note of its respective size, morphology (shape and margin), density, associated calcification, associated features and the location of the mass.
- **Calcifications:** are small deposits of calcium within the breast tissue. When the calcium deposit is coarse, it is called a **Macrocalcification**. They usually are related with signs of age, old injuries or inflammations. They may also be an early sign of cancer when several macrocalcification are found together in one area (form a cluster). When the calcium deposits are lesser than 0.5mm they are called **Microcalcifications**, when many microcalcifications form a cluster in one area, they may indicate a small cancer. Although being a common sign of breast cancer, in many cases microcalcifications are benign. For calcifications it is important to annotate their distribution, associated features and their location.
- **Architectural Distortion:** is a distortion in the breast structure, usually containing calcifications. For architectural distortion it is important to annotate

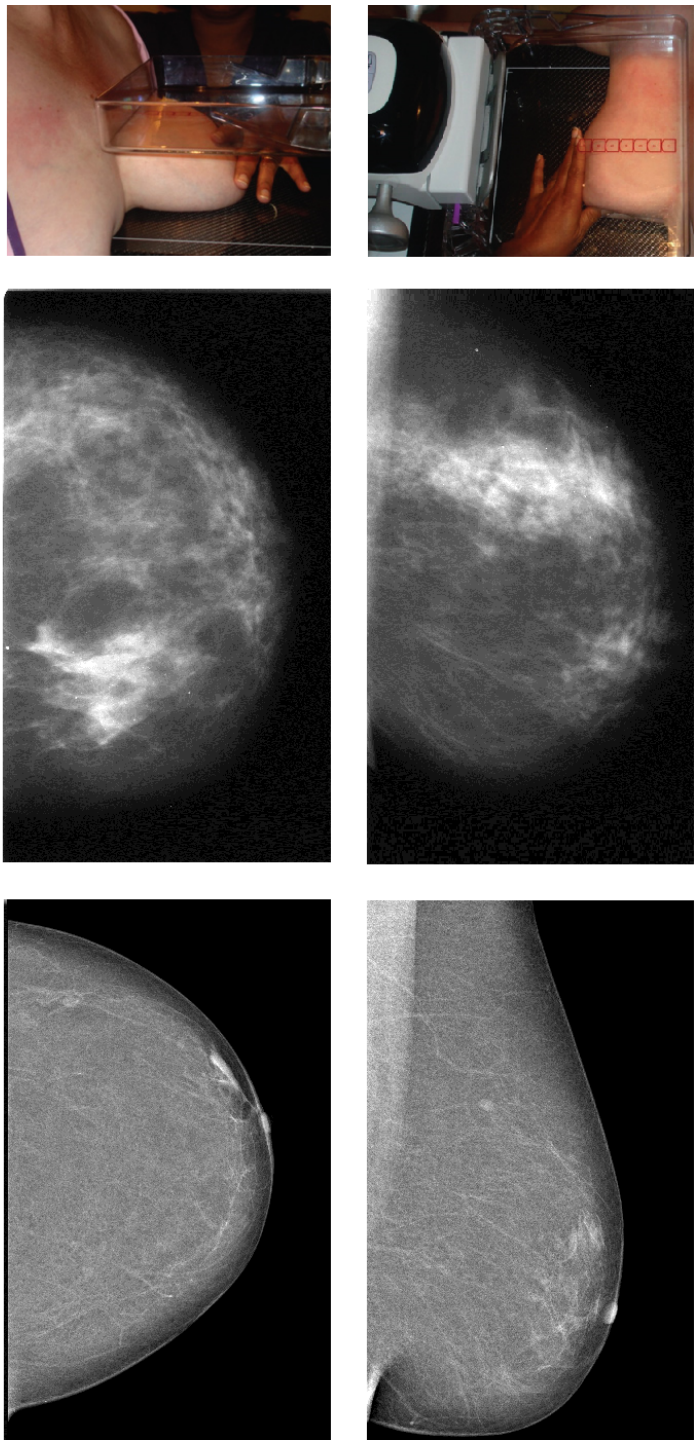


Figure 2.5: Shows on how each view is taken (top), how it is displayed in a mammogram a typical screen-film mammogram (middle) and how its displayed in a digital mammogram. With cranio-caudal view on the left and mediolateral-oblique view on the right. These images were taken from the Inbreast Dataset and from [2].

associated calcifications, associated features and their location.

- **Asymmetries:** Asymmetries in the breast can be classified as global asymmetry when they are asymmetric breast tissue or are greater volumes of breast tissue in comparison with the corresponding region of the opposite breast. Asymmetries can be also classified and focal asymmetry when an area of tissue is visible in two different views, having a similar shape on both views but it does not have the borders of a mass. It is important to annotate the associated calcifications, associated features and the location of this finding.

- **Intramammary lymph nodes:** Are usually nodes with less than 1cm of diameter. They are seen as a circumscribed oval or non-calcified mass with a central or peripheral lucency that represents fat.

- **Skin lesion:** Is a superficial lesion on the breast skin. It is important to annotate the location of this finding to ensure that it is not mistaken for breast lesion.

- **Solitary dilated duct:** Is a dilated duct that may or may not contain calcifications, they have a tubular, slightly nodular shape. Despite being rare, it should not be overlooked, it is the only finding related with malignancy. It is important to annotate the location of this finding.

For sake of visual clarification on how these anomalies look like, each anomaly is represented in Figure 2.6. For more information regarding these findings and how they are related with malignancy consult the breast imaging book[35].

2.4.3 BI-RADS

With the objective of creating a standard classification procedure in mammography the American College of Radiology (ACR) created the BI-RADS classifications and management recommendations presented in the Table 2.2. The ACR also categorizes the breast density in 4 classes illustrated in Figure 2.7, since the higher the density the harder it is to detect breast cancer. A mammogram can be classified in 6 categories:

1. Assessment incomplete where more studies are required regarding the mammogram; Negative where no new findings were detected;

2. Benign finding where the finding was detected and classified as benign;

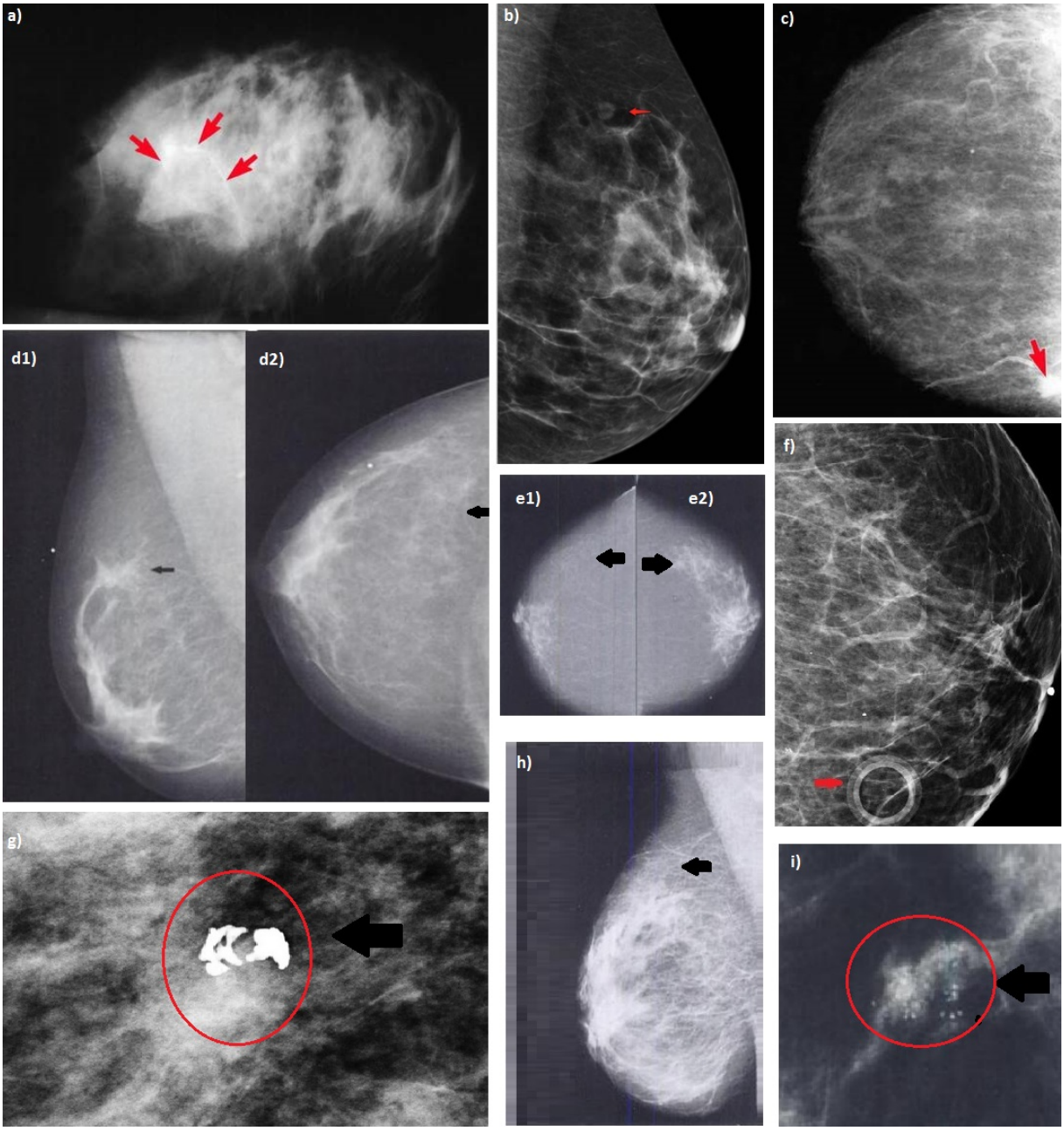


Figure 2.6: Contain several anomalies that can be found in mammogram views such as MLO and CC. These findings are: a)Calcification; b)Intramammary lymph node; c)Malignant mass; d)Focal asymmetry; e)Global asymmetry; f)Skin lesion; g)Macrocalcifications; h)Solitary dilated duct; i)Cluster of microcalcifications. These images were taken from the Atlas of Mammography [3]

3. Probably benign finding where the finding might be benign, a follow up scanning is required to ensure no changes regarding the finding (instead of biennial mammography, the patient should go in 6 months, then 12, then 24);

Table 2.2: Clinical Management Recommendations for Mammograms by Breast Imaging Reporting and Data System (BI-RADS) Category

BI-RADS Category	Assessment	Clinical Management Recommendation(s)	Strength of Recommendation
0	Assessment incomplete	Need to review prior studies and/or complete additional imaging	A
1	Negative	Continue routine screening	A
2	Benign finding	Continue routine screening	A
3	Probably benign finding	Short-term follow-up mammogram at 6 months, then every 6 to 12 months for 1 to 2 years	B
4	Suspicious abnormality	Perform biopsy, preferably needle biopsy	A
5	Highly suspicious of malignancy; appropriate action should be taken.	Biopsy and treatment, as necessary.	A
6	Known biopsy-proven malignancy, treatment pending	Assure that treatment is completed	

4. Suspicious abnormality, usually a biopsy of the finding is required;

5. Highly suspicious of malignancy where biopsy and a treatment if appropriate is required as soon as possible;

6. Biopsy proven malignancy, where only the treatment is pending.

The goal behind this classification and guidelines is to reduce the patient stress and hospital costs by reducing overdiagnosis and overtreatment without reducing the mammography results. For more information check the lexicon created by ACR regarding BI-RADS [36]. In the next chapter we will study the current works to automatize the classification step (between benign and malignant) of anomalies in the mammography exam.

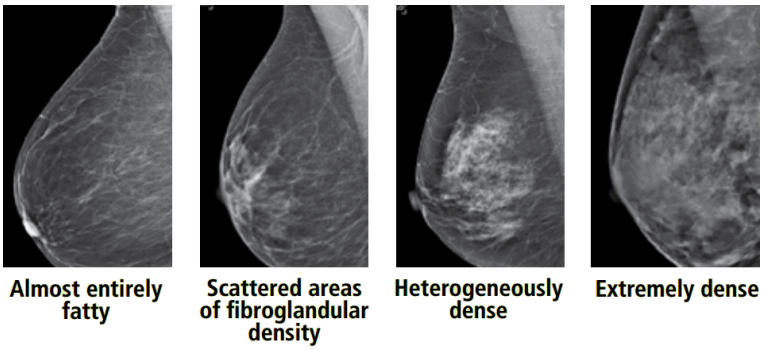


Figure 2.7: The 4 categories of breast density defined by the American College of Radiology from the lowest density type to the highest density type

Chapter 3

Computer Aided Diagnosis for Breast Cancer Classification

3.1 Overview

In the previous chapter, we discussed the main theoretical background on computer vision and our domain of interest: breast cancer. In this chapter we will describe and analyse the state of the art of computer aided breast cancer classification. For most of the studies, a breast cancer classification system is based on the following architecture illustrated in Figure 3.1. First, the user selects a database. This database may contain raw images of mammograms, marked areas of breast anomalies and their respective classification, i.e., which anomaly it is and if it is benign or malignant, clinical information regarding each image and may even contain already processed features from other studies. After selecting a database the next step is to extract image features that might discriminate the anomaly regarding its malignancy. And at last the step where they train a classifier with the database generated features to classify automatically breast cancer. Usually studies in this area focus on one of the previously mentioned 4 active areas of research: First the **image anomaly detection** which consists in using image processing in order to detect anomalies in the breast, a review of works in this step can be found at [37]. However we will not cover this step because its outside the scope of this work, since we will need ground truth anomalies, i.e. marked anomalies by radiologists in order to study their malignancy. We also will not address works based solely on micro calcification classification because they are based on detection of each micro calcification and in the analysis of their distribution

which is outside of the scope of our work. Then the second is the area of **creating a standard mammography database** for the study. This step is still an ongoing research topic since despite the vast amount of available databases of mammograms, since there is no standard database for studies, they all have weaknesses, may that be precise segmentation of the anomaly, or clinical data known related to the anomaly that was not annotated or lack of follow up mammograms among other issues. We will address these issues below at Subsection 3.2. Followed by the third area which is the area of **image pattern retrieval** from the anomalies that may associate the anomaly malignancy, there are a vast amount of experiments in this area, we will analyse and describe them at Subsection 3.3. At last, the fourth area is **breast cancer malignancy classification** where custom made classifiers are designed or common known classifiers are tested with the breast cancer data available by the previous two mentioned areas, breast cancer databases and breast cancer patterns in order to discover which classifier will perform better. Despite each area being addressed separately they rely on each of the remaining, i.e. new databases use image patterns and classifiers to compare their improvement and potential among older databases, image pattern relies on databases to obtain data and classifiers to evaluate their value and classifiers rely on the previous two in order to test their performance. We describe the current works on computer aided breast cancer classification on the three following subsections.

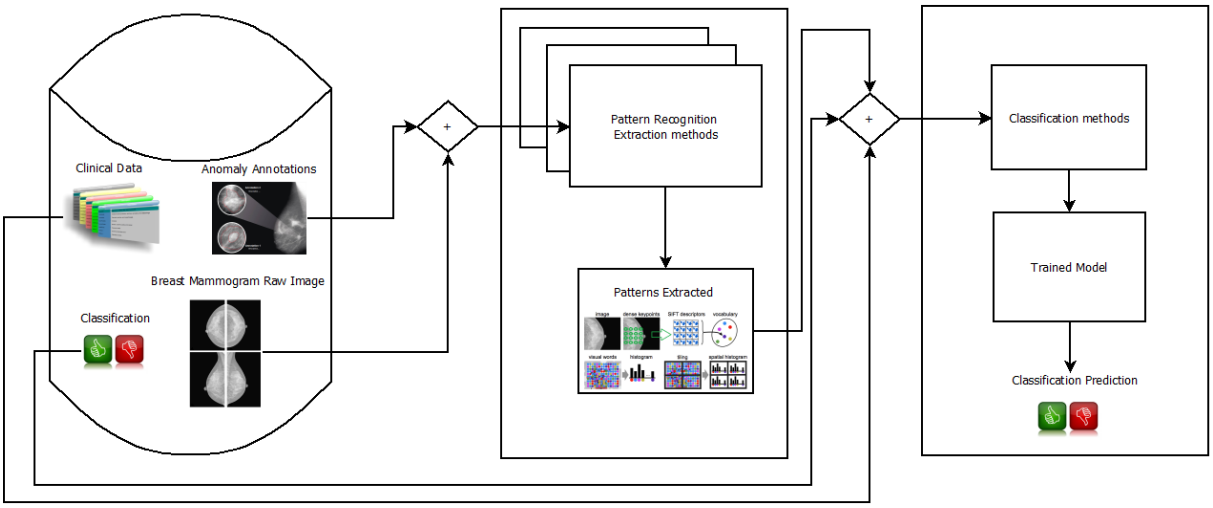


Figure 3.1: The general architecture of a computer aided diagnosis for breast cancer classification. Mammogram images and annotations are used in order to generate image descriptors/patterns that are may be used along with clinical data from the database to construct a classifier that will be able to classify between malignant or benign.

3.2 Breast cancer databases

Breast cancer databases are sets of mammograms images with annotations. There are several databases for breast cancer, some are public such as [38, 39, 40, 41, 42, 43, 44, 45] and others are not [46]. Despite the fact that all those databases contain mammograms as raw base data, their difference lies in the number of cases, availability of biopsy-proven (ground truth classification values), quality of the mammogram and on their annotations. Each database was designed to complete design flaws from existing databases. A good design for a breast cancer database should be a database containing many mammograms with the four standard views (MLO and CC of each breast) and full annotations for each view. These annotations should have the corresponding to the pixel level contour of each finding, the type of finding, density of the breast, BI-RADS classification of the finding and its respective biopsy proven regarding malignancy.

3.2.1 Analogic screening mammography databases

The following databases described are analogic mammograms database. They are useful for overall mammogram breast cancer classification, since the full digital mammography equipment is expensive and their integration at hospitals is still an ongoing process. Many hospitals still use analogic screen film mammography mainly due to lack of funds.

- **Mammographic Image Analysis Society (MIAS)** [38] The MIAS database was one of the first databases available for the public, designed to help the development of computer vision systems to replace human operators of breast cancer detection and also to encourage the creation of more public datasets for the same purpose. The mammograms were taken by the United Kingdom National Breast Screening Programme, containing MLO views with spatial resolution of 50um (microns) and pixel edge taken by a Joyce-Loebl scanning microdensitometer sited at the Royal Marsden Hospital, with 8 bits representing each pixel. There are 4 image sizes small medium large and extra large from 1600 to 5200 pixels x 4320 pixels stored in PGM format. In total, this database contain 322 films and for each film there is the information of the category of the breast density between Fatty, Fatty-glandular and Dense-glandular, the class of the abnormality between calcification, circumscribed masses, spiculated masses,

ill defined masses, architectural distortion, asymmetry and normal, the severity of the abnormality classified as benign or malignant, the coordinates of the centre of abnormality and the approximate radius of the abnormality. This database design fail to give the precise contour of each finding, clinical data regarding each patient of each anomaly and does not contain the BI-RADS annotation of each finding. Despite having this design problem, this database was widely used by the computer vision research community, because it can be instantly accessed via their website and also because it was one of the first datasets for breast cancer detection studies.

- **Digital Database for Screening Mammography (DDSM)** [39, 40] This dataset was created to ease the creation and evaluation of computer aided systems containing 2620 screen mammograms with ground truth and other info completed in 1999. Mammograms obtained from Massachusetts general Hospital, Wake Forest University School of Medicine, Sacred Heart Hospital and Washington University of St. Louis School of Medicine containing the standard four view (left and right MLO and CC views). The spatial resolution and contrast resolution vary according to the hospital equipment between 42um, 43.5um and 50 of special resolution and between 12 and 16 bits per pixel of contrast. The cases were all from mammography exams conducted between a period of 5 months. Each case was digitized and categorized according to 4 categories of BI-RADS (Negative, Benign finding, Probably benign finding and known biopsy-proven malignancy). All cases were automatically cropped to remove as much background (non tissue) as possible. And manually censored any id information of the patients, storing each case in LossLessJPEG format. Each view contain the data of the exam, the ACR breast density. With the exception of two Negative cases, all cases have pixel level ground truth marking of the abnormalies and a description created by a radiologist expert with BI-RADS lexicon. This database for a long time was the closest to a perfect design database i.e. containing patient age, standard ACR breast density classification, pixel level mark of the findings and BI-RADS annotation of each finding. The majority of CAD for breast cancer studies use this database in order to compare performance with previous works in the area. After a wild variety of research performed with this dataset, the only disadvantages found using this dataset were, few studies unrelated to classification reporting lack of precision in their pixel level marking [47, 48], the non existence of clinical data regarding the patient history, the lack of update given the rise of new technologies such as Full Digital Mammograms and the lack of access to the dataset since the current

software used to access the data is deprecated and unavailable. Nevertheless this database is still available with easy access thanks to the IRMA project that will be described below.

- **Image Retrieval in Medical Applications (IRMA) Project Database** [41, 42] The IRMA project is specialized in image retrieval. Taking into account the difficulties of access on the DDSM dataset, they decided to transform the images into a user friendly format (PNG) and to distribute via request at their project site [41]. But they did not stop there, after the experiment of changing they type of the DDSM they concluded that most available databases have different data formats and styles (e.g. Mias PGM, DDSM LLJPEG, LLDL DICOM format) which caused a great trouble for researchers to adapt their works to use all databases. Therefore they decided to create a new database with finding patches from all databases in a standard format. The IRMA Patches database [42] contain patches extracted from DDSM, MIAS, LLNL, and RWTH mammogram databases, resized to 128 x 128 pixel and grouped by their metadata annotation. However, since pre processed patches stray away from the common film digitalized mammogram with segmentation and description of the findings, most researchers still opt for the DDSM dataset as standard database to perform their studies.

- **Bancoweb LAPIMO database** [45] This is the most recent database regarding conventional screening mammography, created on Brazil in 2010. Containing 1400 images from around 320 patients with the four standard views. Each screen film was stored in TIFF format with a contrast resolution of 12 bits per pixel and a spatial resolution varying between 75um to 150um. Including BI-RADS classification, background patient information and breast density classification (not ACR standard). Not all findings are marked, but all findings are described in text. This dataset can be accessed for free via request at their website [45]. This dataset contain several design flaws, such as not all findings are marked in the image, does not follow ACR breast density annotation and despite being a recent database, does not have FFDM cases as the following recent database we will describe.

3.2.2 Digital mammography databases

The following databases described are recent databases that are using full-field digital mammograms. They are useful for overall mammogram breast cancer classification and also displayed an improvement in abnomaly detection and classification in comparison with analogic screen mammography [32]. Due their benefit in cancer detection we believe in the future FFDM will be the standard in the hospitals.

- **Inbreast** [43] A database acquired at Centro Hospitalar de São João, Porto, Portugal. This database was designed to help the current research studies on CAD for cancer detection by covering the design flaws the previous databases had. They used the MammoNovation Siemens FFDM, with a solid-state detector of amorphous selenium pixel size of 70um, 14 bit contrast, 3328x4084 or 2560x3328 resolution according to the plate used in the mammogram aquisition (according to breast size) stored in the DICOM format. Containing 115 cases, where 90 cases contain the standard 4 views and the remaining 25 containing 2 views due the fact they are cases of mastectomy. A total of 410 images. Each image contain annotations made by a specialist in the field also validated by a second specialist, between April 2010 and December 2010, containing breast density acording to ACR, clinical annotations and BI-RADS classification. When there was a disagreement between the experts, the case was discussed until a consensus was obtained. Each case also contain detailed contour and description of their findings. Resulting in a high quality FFDM images with proper annotations and segmentation of the findings.

- **Breast Cancer Digital Repository (BCDR)** [44] This database also was developed at Centro Hospitalar de São João in 2013, since it is really recent it does not have a publication describing it, instead it contains a publication using this dataset to evaluate image descriptors [1]. This database contain four datasets, The BCDR-D01 and BCDR-D02 containing FFDM images of 230 biopsy-proven lesions of l79 patients and 162 biopsy-proven lesions of 64 patients respectively. The remaining two datasets BCDR-F01 and BCDR-F02 have analogic screening mammography images containing 200 biopsy-proven lesions of 190 patients and 188 biopsy-proven lesions of 98 patients. All datasets contain clinical data regarding the patient, the image descriptors extracted at [1], the segmentation of each and description lesion and the patient binary classification between benign and malignant.

- **Others** There are still other databases which are not available for the public (Paid databases or Private Registration required), e.g. [46] and Lawrence Livermore National Laboratory and University of California San Francisco Database[49]. The Created by L. L. N. Laboratories and University of California at San Francisco Radiology Department, the LLNL/UCSF database uses films digitized to 35 microns where ach pixel was sampled to 12 bits of grayscale. Containing a total of 198 films with 4 views from 50 patients (except those with mastectomy containing only 2 views) separated according to 4 categories of BI-RADS containing 5 Negative cases, 5 Negative but difficult cases (with either dense or fibrous breasts, implants, or asymmetric tissue), 20 cases of Probably Benign microcalcifications (with at least 3 years of follow-up without change or developing cancer), 12 cases of suspicious abnormality benign microcalcifications, (note: all these benign cases had either a biopsy or a diagnostic mammogram plus at least 3 years of subsequent follow-up without change or developing cancer), and 8 cases with a biopsy-proven malignancy of microcalcifications, according to [50] available at the price of 100 US $ to cover their reproduction costs. It also contains pixel masks indicating where are the findings. However it still contain design flaws such as no clinical associated data. Usually these databases also are not common in the research community, because most researchers do not have access to them. Since most of their information are not publicly available (such as other databases described with unknown information at [43] that we do not mention), we will not add these databases on the following summary of all the databases previously described (with the exception of LLNL UCSF database because it can be currently accessed without payment at their original website [49]). For the sake of summarizing we created the table 3.2.2, containing all the information of all the previously discussed databases. The most used databases in the literature are the MIAS and the DDSM database due their easy access, although the MIAS usage has been decreasing along the years since it does not have BI-RADS classification, BI-RADS annotation and the precise leisure contour. It is also important to remark novel databases such as BDCR and inBREAST because of the new different data they contain such as FFDM mammograms and full description of the findings according to BI-RADS lexicon and also surpass a good database design such as DDSM which may lead new studies of FFDM mammograms.

Databases	MIAS[38]	Bancoweb[45]	DDSM(IRMA)[41]	LLNL/UCSF[49]	inBREAST[43]	BCDR-D[44]	BCDR-F[44]
Year	1994	2010	1999 Unknown	2010	2013	2013	
Origin	UK	Brazil	USA	USA	Portugal	Portugal	Portugal
Number of Cases	161	320	2620	50	115	243	288
Number of Images	322	1400	10480	198	410	392	388
Views	MLO	CC MLO	CC MLO	CC MLO	CC MLO	CC MLO	CC MLO
Contain Mastectomy	No	Unknown	No	Yes	Yes	No	No
Mammogram Type	Screen Film	Screen Film	Screen Film	Screen Film	FFDM	FFDM	Screen Film
Contrast Resolution	8	12	12 and 16	12	8	8	Unknown
Image Format	PGM	TIFF	PNG	DICOM	DICOM	TIF	TIF
Patient data	No	Yes	Age	Yes	Yes	Yes	Yes
Abnormally Categorization	Yes	Yes	Yes	Calc. Only	Yes	Yes	Yes
Clinical data	No	Yes	Age	Yes	Yes	Yes	Yes
Abnormally Annotation	No	Yes	BI-RADS	Yes	BI-RADS	Yes	Yes
Classification type	Binary	BI-RADS	BI-RADS	BI-RADS	BI-RADS	Binary	Binary
Abnormally Contour	Radius	Few cases	Yes	Yes	Yes	Yes	Yes
Image Descriptors	No	No	No	No	No	Yes	Yes

3.3 Image Pattern Retrieval

In this section we address the current developed patterns for breast cancer images. But before extracting features, according to Mohanty et al. [12], pre-processing is essential in order to improve the quality of the mammogram, they explain which techniques are used to remove noise, enhance structures and enhance contrast. Depending on the anomaly, extra information may be contained for example, at Moreira et al. [51], they describe intensity patterns and shape patterns extracted from masses using the InBREAST database, and make an interesting study regarding their relationship with malignancy. In some works they use clustering information regarding micro calcifications [52, 53, 54] plus other image patterns to conclude the cluster relationship with malignancy. Nevertheless, in general, intensity patterns and other types of image patterns can be extracted from any anomaly and used for cancer classification, we will address these works in more detail below.

3.3.1 Image Patterns

After analysing the current state of the art initially based on surveys regarding computer aided diagnosis for breast cancer such as [9, 10, 1], we were able to separate on the following image pattern types.

- **Grey Level Co-occurrence Matrices (GLCM) and intensity statistics** are well known for their performance in texture analysis and simplicity, therefore several features regarding GLCM and intensity histograms were studied in the literature, such as gray-level correlation, entropy and roughness in [51], or in

[55] that analyses 16 statistical features and 21 GLCM based texture features (some of the studied features: mean, deviation, smoothness, skewness, energy, dissimilarity, difference, length, among others). [56] also uses several of the intensity descriptors previously mentioned and also introduce the local binary patterns in this problem, where local binary patterns outperformed the other statistics features being reported as a good descriptor for breast cancer classification. There are also reports [57, 54, 58] that indicate the good performance of Haralick features (obtained from GLCM).

- **Multi scale methods** are widely used for breast cancer classification. The most popular multi scale method is the wavelet which can be adapted by using different mother wavelet functions (as we saw in the background). Several wavelet functions were studied for breast cancer anomaly classification, such as Daubechies [9, 59, 60, 61, 53, 54], symlet [9], bi-orthogonal [9], Haar [9, 61, 53], Gabor [57, 62] and others [52, 53, 63, 54]. There are other multi resolution methods based on the wavelets explored for this purpose such as Curvelets that explore the curve instead of the wave function [9, 64], and also the Ridgelet method [63] which was explored recently, howerver, the results did not outperform the wavelet method. In order to compare these multi resolution methods, Ramos *et al.* [9] did a comparison work studying several wavelet functions against Curvelets, concluding that Curvelets outperform the tested wavelets methods for breast cancer anomaly classification.

- **Local invariant features** were not popular for breast cancer classification, however, Moura *et al.* [1] proposed to use local invariant features to classify masses since they contain great information regarding shape. One of their experiments was to use a Histogram of Oriented Gradients (HOG) [65] based on the famous Scale Invariant Feature Transform [16], which divide the ROI in a grid of blocks and for each block calculate a histogram of the orientation of the gradient. After experimenting HOG they proposed a new image descriptor to describe the regularity of the masses shapes in breast images, this method was called the Histogram of Gradient Divergence (HGD) [1]. Based on the principle that gradient of boundaries pointing to the centre of the object is a characteristic of round-shaped objects with continuous regular border, they assume that the object is in the center of each patch and measure the gradient divergence of a pixel as the angle between the vector of the intensity gradient on the pixel and a vector with origin on the pixel pointing to the center of the patch. Rotation invariance is obtained naturally in this method since they store the divergence of

the gradient instead of the orientation of the gradient. With this novel descriptor they outperformed all previous image descriptors for mass classification.

3.3.2 Discussion

Given the previously mentioned methods for extracting features from a region of interest, we can conclude that these features are heavily dependent of the type of scanner. Meaning, features extracted from a type of scanner will perform worse on other scanners [10], also they contain interesting information regarding malignancy where these works mentioned review a high classification performance. Apparently the wavelet, curvelet and HGD features seems to be the best methods for pattern retrieval, given its good results in comparison with other methods stated in their respective works. Since these methods are all using different databases or types of classifiers or types of features to diagnose (e.g. microcalcifications clusters, calcifications, all anomalies, etc) we cannot create a table summarizing their performance because we cannot directly compare them. Fortunately, Moura et al. [1], compiled most of these results and analysed them using two datasets (DDSM and BCDR), separating by masses and microcalcifications using a linear SVM classifier to classify given each feature between benign or malignant using the Median Area Under Curve (AUC) as performance metric. This study contradict the Curvelet superior performance against wavelets from [9] as we can see in the Table 3.1 where we can compare most of the methods and their performance, with also the addition of clinical data available by the BCDR and DDSM datasets. We can conclude that the most discriminant pattern regarding all abnormalities or only masses is the HGD while the most discriminant patterns for calcifications are the Gabor filter, the Wavelets and the Haralick features.

3.4 Machine learning

In this section we will address the machine learning methods used to classify and evaluate the previously discussed breast cancer image features. During this research we noticed that most works use the SVM method [1, 57, 62, 66, 67] or the kNN method [9, 61, 64, 58] for this purpose. We believe that the preference for these two methods is based on the following reasons. The kNN is quite common on image processing due to the matching nature of the algorithm (explained on the previous chapter). The SVM

Table 3.1: Classification performance (AUC) of the standalone clinical data and of the image descriptors (standalone and combined with clinical data) from [1]

Data set	Standalone clinical data	Combined with clinical data	Image descriptors											
			IS	HM	IM	Zer	Har	GLRL	GLDM	Gab	Wav	Curv	HOG	HGD
All Lesions														
DDSM sample	0.853	No	0.715	0.691	0.667	0.691	0.736	**0.743**	0.683	0.725	0.731	0.712	0.729	0.736
		Yes	**0.868**	0.860	0.859	0.864	0.857	0.862	0.845	0.854	0.860	0.865	0.848	0.851
BCDR F01	0.712	No	0.637	0.614	0.691	0.648	0.710	0.654	0.641	0.712	0.719	0.705	0.739	**0.825**
		Yes	0.766	0.765	0.770	0.754	0.784	0.713	0.743	0.788	0.776	0.781	0.765	**0.817**
Masses														
DDSM sample	0.867	No	0.707	0.667	0.647	0.675	0.718	**0.733**	0.683	0.711	0.720	0.703	0.707	0.732
		Yes	**0.890**	0.882	0.887	**0.890**	0.885	0.879	0.880	0.878	0.884	0.887	0.877	0.883
BCDR F01	0.829	No	0.670	0.648	0.681	0.740	0.765	0.688	0.695	0.764	0.768	0.712	0.788	**0.860**
		Yes	0.844	0.830	0.841	0.833	0.876	0.799	0.823	0.848	0.849	0.843	0.841	**0.894**
Calcifications														
DDSM sample	0.807	No	0.733	0.754	0.700	0.718	**0.774**	0.764	0.695	0.766	0.773	0.729	0.717	0.706
		Yes	0.799	0.779	0.787	0.791	0.797	0.787	0.769	**0.803**	0.792	0.783	0.764	0.777
BCDR F01	0.725	No	0.711	0.704	0.728	0.617	**0.793**	0.694	0.683	0.790	0.765	0.756	0.710	0.778
		Yes	0.790	0.768	0.783	0.741	**0.815**	0.737	0.728	**0.815**	0.801	0.800	0.747	0.783
Number of wins			2	0	0	1	3	2	0	3	2	0	0	**8**

The highest score of each scenario is highlighted at bold, and scores with no evidence of differences to the highest ($p < 0.05$) are underlined. The last row shows the total number of times each descriptor achieved the highest (or comparable to highest) score
IS intensity statistics, *HM* histogram measures, *IM* invariant moments, *Zer* Zernike moments, *Har* Haralick features, *Gab* Gabor filter banks, *Wav* wavelets, *Curv* curvelets

is the second favourite due to its good performance sparse and noisy data. Despite the effectiveness of both methods, other classifier approaches such as decision trees [68, 63], neural networks [69, 60], rule based [70, 58] were used in order to obtain study their performance boost on this purpose. We will review the current methods used for breast cancer classification, reporting their performances.

3.4.1 kNN based methods

As we saw in the previous chapter, this method uses a dataset as the whole classification model where new data is classified by the class of the nearest data entry on the database given by a proximity function. This method was used by [9] to evaluate features extracted from the Mias dataset, obtaining an accuracy of 100% to classify benign cases and 83,3% accuracy to classify malignant cases. Another work [61], also extract features from the Mias dataset obtaining an accuracy of 98,8% to classify between malignant or benign. There also another approach used by [58], which combines the kNN with a rule based method, i.e. first he used rule based methods to transform the extracted features into rules and then used the kNN to match the rules, obtaining 90% accuracy also using the features from the Mias dataset.

3.4.2 SVM based methods

In a very brief description, the SVM method aims to find the best hyperplane to split the classes from a dataset. We will address how this hyperplane is calculated in the next chapter. Several works reported good performance using the SVM classifier, such as [1] obtaining an AUC of 0.89 and [57] which obtained an accuracy of 91.4%, where both used features extracted from the DDSM dataset. Another work [66] used SVM on features extracted from the Mias dataset obtaining an AUC of 0.91. Besides the direct usage of the SVM method, other works attempted to adapt the SVM in order to obtain better results with breast image data. In [62], a method called "proximal SVM" is introduced, obtaining an AUC of 0.78 on the Mias data, where the author claim that a proximal SVM approach perform better than the normal SVM for breast cancer classification. Other work [67], which used the DDSM dataset to extract features, realized that depending on the image feature extracted, the kernel used by a SVM to fit the data might change. Therefore, they proposed to create several SVMs with different kernels for each set of data and ensemble them in a voting system. Where the system would select the result based on the majority of all SVMs results, achieving an improvement of 0.02 AUC over the SVM method. Obtaining a final AUC of 0.92. This majority vote SVM strategy resemble the idea of the MKL method that we will explain on the next chapter, where we will also talk about kernels and how MKL use them to turn the SVM into a more flexible method.

3.4.3 Other methods

There are also other approaches used in the literature. Decision Tree methods are explored by [68], where in their work they compare the results between Simple CART, Random Tree and Random Forest, obtaining accuracy of 96.5% in their best result (Random Forest) using the Wisconsin dataset. Another work [63] also uses Random Forest, but they first use a Genetic algorithm to filter the features obtained from the DDSM dataset and then they use the filtered features to create a Random Forest classifier, they obtained an AUC of 0.90. Besides decision tree based methods, there are other approaches using Neural Networks. In [69] they analyse several types of Neural Networks methods. Such as Back Propagation Neural Networks, Radial Basis Function Network, Modular Neural Networks and Artificial Neural Network. They obtained their best result from the Modular Neural Network (98.2% accuracy), using the Wisconsin dataset. There are also two other works that explore Neural Networks.

The [60] where they use a Neuro Fuzzy Logic classifier obtaining an accuracy of 93.7% on the Mias dataset. And the work [70], where they explain the weaknesses of neural networks and rule based methods and create a method to overcome them by combining both methods, obtaining a sensitivity of 100% and a specificity of 69.2%.

3.5 Summary

After analysing all those areas and studies we can remark the following. There are many works for masses and calcifications. However, there is still room for improvements by experimenting novel techniques of classification or by uncovering new image patterns. There is also little study regarding other findings such as architectural distortion and asymmetry. There is a big variety of effective image descriptors for breast cancer classification that can be selected. But, there are not many approaches on taking full advantage of these image features on the machine learning side, i.e. more works that attempt new strategies to increase learning from image features. Such works as the majority vote SVM[67], the rule base methods combined with kNN[58] or Neural Networks[70] are attempting to improve the performance by using more adaptable approaches for the breast cancer data new instead of using old classifier methods without adapting them to the problem. It is hard to summarize the performance results obtained with the reported classification methods since each work vary on the usage of the database, or in the patterns extracted from each database or even in the metric used to display the results. Nevertheless, it is possible to conclude that classifiers adapted for the breast cancer image data can obtain better results. Given that fact we are motivated to use machine learning methods that were not explored yet for the breast cancer classification problem, such as the Multiple Kernel Learning (MKL) which will be addressed on the next section. We also noticed a high difficulty and ambiguity on comparing results with different works. Therefore we propose a comparison study to compare the usage of MKL against one of the most common machine learning method used, the SVM.

Chapter 4

Multiple Kernel Learning

After reviewing the literature on machine learning methods for breast cancer, we found several methods with good performance using kernel methods such as SVM but we did not find any work regarding multiple kernel learning for this purpose. Therefore in this chapter we will describe in depth the single kernel method SVM at Section 4.1, then in Section 4.2 we will see how the multiple kernel learning extend the single kernel method, increasing the model flexibility toward the training data in general, then we describe the MKL algorithm we selected for our study, the Simple MKL [8]. In the end we summarize the MKL method focusing on the advantages that we expect to achieve by applying it in breast cancer classification. The sections 4.2 and 4.3 were based on the Simple MKL publication [8].

4.1 Introduction

Multiple kernel learning is based on single kernel methods such as Support Vector Machines (SVM). In the Background chapter we did not explain in depth how SVM works because it will be simpler for the reader to learn the MKL extension of the SVM method after understanding how the maximum margin hyperplane problem is formulated below. In depth, the linear SVM solves the maximum margin hyperplane search problem, which can be formulated as it follows. Given a training data $\boldsymbol{D}$ with a set of n points of the Eq. (4.1) where the y_i is either 1 or -1 (other classes). In order to split a set of x points in $\boldsymbol{D}$ its required to calculate a hyperplane that split these points, which is represented by Eq. (4.2). When the Eq. (4.2) equals 1 or -1, it means that the hyperplane can be used as margin. Because it separate the classes from the

feature set. Points laying on the margins are called "support vectors" since they are the vectors that form the margins. The distance between two margins is given by $\frac{2}{\|w\|}$. In order to obtain the maximum margin hyperplane we minimize the norm of w. However, this problem depends on the norm of w which involves a square root. In order to optimize it, Lagrange multipliers were introduced, transforming the problem into a dual maximization problem [20, Ch. 7, p. 325].

$$D = \{(x_i, y_i)|x_i \epsilon \mathbb{R}^p, y_i \epsilon \{-1, 1\}\}_{i=1}^{n} \tag{4.1}$$

$$\mathbf{w} \cdot \mathbf{x} - b = 0 \tag{4.2}$$

When the problem cannot be solved linearly, i.e. there is no linear hyperplane that can separate the classes from the given dataset, the feature data space is mapped to a higher dimension space where exists a hyperplane that can split the dataset classes, an example is illustrated at Figure 4.1. This mapping is performed by using Kernel functions, adapting from linear functions to non-linear functions such as Polynomial 4.3, Gaussian 4.4, Radial 4.5 and others. With their respective kernel parameters such as degree (d), Gaussian and radial width (σ, γ). Thanks to the "kernel trick" [71] there is no need to compute the dot products in these high dimensional, since it allows to compute the dot products within the original feature space by the means of a kernel function.

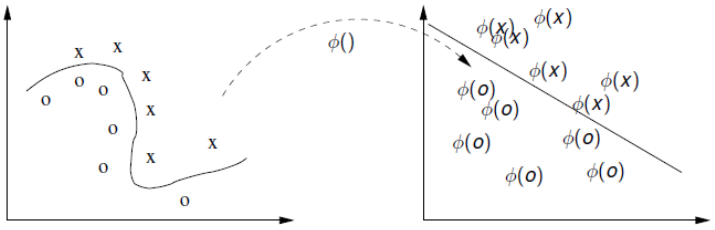


Figure 4.1: The mapping between the data in its original space and the data transformed into a higher dimension space by the usage of an appropriated kernel function where the data can be easily separated according to their classes.

$$K(x_i, x_j) = x_i \cdot x_j^d \tag{4.3}$$

$$K(x_i, x_j) = e^{(-\frac{\|x_i - x_j\|^2}{2\sigma^2})} \tag{4.4}$$

$$K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2} \tag{4.5}$$

Sometimes, the data may have mislabelled classes or outliers that may shorten the margin, in order to adapt to these cases, slack variables were introduced by [25]. These slack variables are added to each training data point allowing to have points outside crossing their class margin with a certain cost penalty which increases linearly. The trade-off between the margin and the slack variables can be defined by a variable C (C>0). In summary, the SVM learning problem can be formulated as Eq. (4.6), where the dot product is replaced by the selected kernel and where α (constrained by C) and b are coefficients to be learned from the train data. Additional information regarding support vector machines can be found at [20].

$$SingleKernel = \sum_{i=1}^{n} \alpha_i K(x, x_i) + b \tag{4.6}$$

4.2 Multiple Kernel Learning model

Kernel methods such as SVMs were proven efficient for classification and regression. However, the SVM uses a single kernel to fit the entire data. As we saw previously, the SVM can be formulated in Eq. (4.6), where K($\cdot$,$\cdot$) is a given positive definite kernel associated with a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$[72]. MKL extends this equation into a more adaptable equation by working with a set of kernels using a combination of weights turning the problem into a convex optimization problem represented by the Eq. (4.7). For example, given M kernels k_1, ..., k_m that are potentially suited for a given problem, the MKL consist in the problem of finding the positive linear combination of these kernels resulting on an "optimal" kernel. In order to give a clear insight regarding the adaptability to data using multiple kernels we created the Figure 4.2 where we suppose that we have 4 kernels (k_1, k_2, k_3, k_4) and only the first kernel and the fourth kernel perform well with a certain data, we adapt the optimal kernel by adding each kernel multiplied by a weight d_{mi} according to their performance, e.g. 0.5 weight for d_{m1} and d_{m4} and 0.0 weight for weight d_{m2} and d_{m3}.

$$MKL(x) = \sum_{i=1}^{n} \alpha_i \sum_{m=1}^{M} d_m k_m(x, x_i) + b \tag{4.7}$$

Before we start explaining how the multiple kernel learning problem is optimized, we will describe the framework that was used as base for simple MKL optimization algorithm. Assume that K_m, m = 1,...,M are M positive definite kernels on the same input space, each of them being associated with an RKHS $\mathcal{H}_m$ endowed with an inner

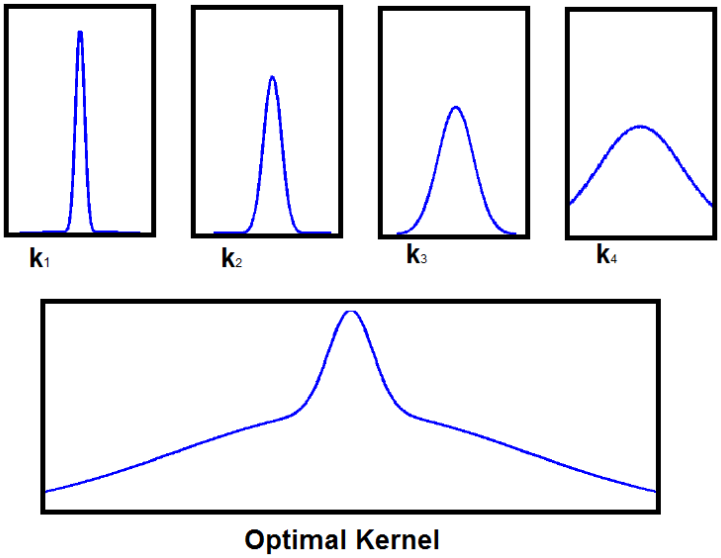


Figure 4.2: The distribution of 4 kernels with different parameters and the sum of all these 4 kernels according to the following weights 0.5 weight for k_1 and k_4 and 0.0 weight for weight k_2 and k_3.

product $\langle\cdot,\cdot\rangle_m$. For any m, let d_m be a non-negative coefficient and $\mathcal{H}'_m$ be the Hilbert space derived from $\mathcal{H}_m$ in Eq. (4.8) endowed with the inner product in Eq. (4.9).

$$\mathcal{H}'_m = \{f | f \in \mathcal{H}_m : \frac{\|f\|\mathcal{H}_m}{d_m} < \infty\} \tag{4.8}$$

$$\langle f, g\rangle_{\mathcal{H}'_m} = \frac{1}{d_m}\langle f, g\rangle_m \tag{4.9}$$

Using the convention that $\frac{x}{0}$=0 if x=0 and ∞ otherwise. This means that, if d_m=0 then a function f belongs to the Hilbert space $\mathcal{H}'_m$ only if f=0 $\in \mathcal{H}_m$. In such a case, $\mathcal{H}'_m$ is restricted to the null element of $\mathcal{H}_m$. Within this framework, $\mathcal{H}'_m$ is a RKHS with kernel K(x,x') = d_mK$_m$(x,x') since $\forall f \in \mathcal{H}'_m \subseteq \mathcal{H}_m$, the decision function is represented by Eq. (4.10). If we define $\mathcal{H}$ as direct sum of the spaces $\mathcal{H}'_m$, then, a classical result on RKHS [72] says that $\mathcal{H}$ is a RKHS of kernel represented in the Eq. (4.11).

$$f(x) = \langle f(\cdot), K_m(x,\cdot)\rangle_m = \frac{1}{d_m}\langle f(\cdot), d_m K_m(x,\cdot)\rangle_m = \langle f(\cdot), d_m K_m(x,\cdot)\rangle_{\mathcal{H}'_m} \tag{4.10}$$

$$K(x,x') = \sum_{m=1}^{M} d_m K_m(x,x') \tag{4.11}$$

Thanks to this simple construction, the simple MKL creators were able to build a RKHS $\mathcal{H}$ for which any function is a sum of functions belonging to $\mathcal{H}_m$. In their framework the MKL aims to determine the set of coefficients $\{\mathrm{d}_m\}$ within the learning process of the decision function. Therefore envisioning the MKL problem as learning a predictor belonging to an adaptive hypothesis space endowed with an adaptive inner product. Thus, the problem of learning the weights d_m, learning the α and b at the same time can be addressed by solving the convex problem referred as primal MKL problem stated below:

$$\min_{\{f_m\},b,\xi,d} \quad \frac{1}{2}\sum_m \frac{1}{d_m}\|f_m\|^2_{H_m} + C\sum_i \xi_i \tag{4.12}$$

$$s.t. \quad y_i \sum_m f_m(x_i) + y_i b \geq 1 - \xi_i \quad \forall i \tag{4.13}$$

$$\xi_i \geq 0 \quad \forall i \tag{4.14}$$

$$\sum_m d_m = 1, \quad d_m \geq 0 \quad \forall m \tag{4.15}$$

Where each d_m controls the squared norm of f_m in the objective function. The smaller the d_m is, the smoother f_m should be.
There are many solutions proposed described in [8], but we will only describe in the following subsection, how the simple mkl algorithm solve the primal problem stated in Eq. (4.12), for other solutions or proofs of the equations described above check [8].

4.3 Simple MKL

In order to solve the problem stated in Eq. (4.12), the Simple MKL formulate the following constrained optimization problem:

$$\min_d J(d) \text{ such that } \sum_{m=1}^{M} d_m = 1, d_m \geq 0 \tag{4.16}$$

where

$$J(d) = \begin{cases} \min\limits_{\{f\},b,\xi} \frac{1}{2}\sum\limits_m \frac{1}{d_m}\|f_m\|^2_{H_m} + C\sum\limits_i \xi_i \quad \forall i \\ \text{s.t.} \quad y_i \sum\limits_m f_m(x_i) + y_i b \geq 1 - \xi_i \\ \xi_i \geq 0 \quad \forall i \end{cases} \tag{4.17}$$

They solve the problem 4.16 on the simplex by a simple gradient method which is addressed below. Knowing that the objective function $J(\mathrm{d})$ is actually an optimal SVM objective value and that the gradient of $J(\cdot)$ can be computed, i.e. the $J(\mathrm{d})$ can be obtained by any single kernel machine, which ties the overall complexity of the SimpleMKL to the complexing of the single kernel machine. For an in depth reading of the intermediate steps on how they managed to differentiate $J(\mathrm{d})$ and compute the gradient $J(\cdot)$ please refer to [8].
Once the gradient of $J(\mathrm{d})$ is computed, d is updated using a descent direction ensuring that the equality constraint and the non-negativity constraints on d are satisfied. The equality constraint is handled by computing the reduced gradient method [73]. The positivity constraints have also to be taken into account in the descent direction. Since they want to minimize $J(\cdot)$, the negative reduced gradient of J is a descent direction. However, if there is an index m such that $\mathrm{d}_m = 0$ and the reduced gradient of J of $_m$ > 0, using this direction would violate the positivity constraint for d_m. Therefore, in these cases the descend direction (D_m) for that component is set to zero. The descend direction update is defined by the Eq. (4.18) where d_μ is a non-zero entry of d.

$$D_m = \begin{cases} 0 & \text{if } \mathrm{d}_m = 0 \text{ and } \frac{\partial J}{\partial d_m} - \frac{\partial J}{\partial d_\mu} > 0 \\ -\frac{\partial J}{\partial d_m} + \frac{\partial J}{\partial d_\mu} & \text{if } \mathrm{d}_m > 0 \text{ and } m \neq \mu \\ \sum\limits_{g \neq \mu, d_v > 0} (\frac{\partial J}{\partial d_v} \frac{\partial J}{\partial d_\mu}) & \text{for m} = \mu \end{cases} \tag{4.18}$$

Now that we know how the descent direction is updated, we can understand the SimpleMKL Algorithm 1. Once the descend direction is computed, we first look for the maximal admissible step size (γ) in that direction and check whether the objective value decreases or not. The maximal admissible step size corresponds to a component d_v, set to zero. If the objective values decreases, d is updated, we set $\mathrm{D}_v = 0$ and normalize D to comply with the equality constraint. Repeating this procedure until the objective stops decreasing. At this point, we look for the optimal step size γ, which is determined by using a one-dimensional line search, with a proper stopping criterion to ensure global convergence. In order to obtain the optimal conditions, this entire procedure is repeated until stopping criterion such as the duality gap (the difference between the primal and dual objective values, where zero is optimal) or KKT conditions 4.19 or the variation of d between two consecutive steps, or even set by a maximum number of iterations.

Algorithm 1 SimpleMKL Algorithm

set $d_m = \frac{1}{M}$ for m = 1,...,M
while stopping criterion not met **do**:
    compute J(d) using a SVM solver with $K = \sum_m d_m K_m$
    compute $\frac{\partial J}{\partial d_m}$, for m = 1,...,M and descent direction D(4.18)
    set $\mu = \underset{m}{argmax} d_m$, $J^\dagger = 0, d^\dagger = d, D^\dagger = D$
    while $J^\dagger < J(d)$ **do** {descent direction update}
        $d = d^\dagger, D = D^\dagger$
        v= $\underset{\{m|D_m<0\}}{argmin} -d_m/D_m, \gamma_{max} = -d_v/D_v$
        $d^\dagger = d + \gamma_{max} D, D^\dagger_\mu = D_\mu - D_v, D^\dagger_\mu = 0$
        compute $J^\dagger$ by using a SVM solver with $K = \sum_m d^\dagger_m K_m$
    end while
    line search along D for $\gamma \in [0, \gamma_{max}]$ {calls an SVM solver for each γ trial value}
    d$\leftarrow d + \gamma D$
end while

$$\frac{\partial J}{\partial d_m} + \lambda - \eta_m = 0 \quad \forall m \tag{4.19}$$

$$\eta_m . d_m = 0 \quad \forall m \tag{4.20}$$

where λ and $\{\eta_m\}$ are respectively the Lagrange multipliers for the equality and inequality constraints of the problem 4.16.

4.4 Conclusion

There are many methods regarding multiple kernel learning in the literature [8], but we selected the simple MKL for two main reasons. First, because the Simple MKL has an open source implementation in Matlab. Second, because its simple to use in comparison with other MKL methods, which is great for our study since our goal is to evaluate the performance of MKL and prove that it can reach higher results than single kernel methods. For sake of summarizing we described the Simple MKL method in Algorithm 2 where first they add equal weights for all kernels, and then the algorithm optimizes these weights by calculating the objective values using a SVM solver, the descent direction and also by calculating the optimal stepsize for that direction, resulting in the update of the weights. The algorithm stop this optimization when the criterion conditions met. After learning in depth how the multiple kernel

learning extend the SVM method in order to adapt the kernel into an optimal kernel for a set of data we can understand why it might surpass the SVM method in terms of performance, i.e., the MKL is more flexible toward the data than the SVM. In order to prove this performance improvement we perform in the next chapter an study using breast cancer image patterns as data in order to classify their malignancy using SVM and Simple MKL.

Algorithm 2 Simplification of the Simple MKL Algorithm

1: set $d_m = \frac{1}{M}$ for m = 1,...,M
2: **while** stopping criterion not met **do**:
3: compute J(d) using a SVM solver with $K = \sum_m d_m K_m$
4: compute $\frac{\partial J}{\partial d_m}$, and projected gradient as descent direction D
5: $\gamma \leftarrow$ compute optimal stepsize
6: d$\leftarrow d + \gamma D$
7: **end while**

Chapter 5

Experimental Study and Results Discussion

In this chapter we describe the conducted tests to evaluate the usage of MKL in breast cancer and discuss the results obtained during this experimental study. The main goals of this study is to prove the existence of a performance difference between the MKL and SVM method and also to analyse the performance of MKL with breast cancer image data. For this study we selected the BCDR-F01 dataset, and extracted from it several image descriptors such as Clinical Data, Intensity Descriptor, Wavelet Descriptor, Local Binary Pattern Descriptors and Histogram Divergence Gradients to use as input data to train, evaluate and compare MKL and SVM classifiers. In more detail, we describe the selection dataset used for evaluation in 5.1, the selection features and how they were extracted 5.2, the MKL method and the kernels that were used in 5.3, the evaluation method for the experimental study 5.4 and the discussion of the results obtained in 5.5.

5.1 Database Description

In order to perform our study to make a comparison between the performance of SVM and MKL toward breast image patterns, we require a breast image database containing clinical data, pixel level contour of the findings on each breast image with their respective biopsy proven malignancy classification. Since the BCDR01-F01 satisfies all those requirements and because it is a local dataset (from Porto) that we had access since the beginning of this work, we selected it to perform our experiments.

This dataset contains screening mammography images of 200 biopsy-proven lesions of 190 patients, described below in more detail.

- **MLO**: Containing 183 images with annotated findings where 116 of the images contain masses, 95 of the findings are benign and 88 of the findings are malignant.
- **CC**: Containing 179 images with annotated findings where 115 of the images contain masses, 92 of the findings are benign and 87 of the findings are malignant.

5.2 Feature Extraction

Since the dataset contains images of the whole breast that was taken during the screening mammography, we required to perform some pre-processing steps in order to obtain images with only the annotated findings. We created a Matlab script to perform the following pre-processing steps illustrated in Figure 5.1:

1. Create a polygon surrounding the finding using the pixel level annotation of the finding.
2. Get the biggest and smallest x and y values of the pixels in the polygon to create a bounding box surrounding the finding.
3. Crop the image using the bounding box calculated.
4. Subtract the intensity value of all pixels outside the polygon created.

After obtaining the pre-processed images from the BCDR-F01. We extracted data descriptors using the image pattern extraction methods (with the exception of the Clinical Data and the Intensity values that were already available in the dataset) described below.

5.2.1 Clinical Data

This data was obtained directly from the dataset. Containing information of the patient and the findings on the image. This subset will be mentioned as *S1* and contains the following features:

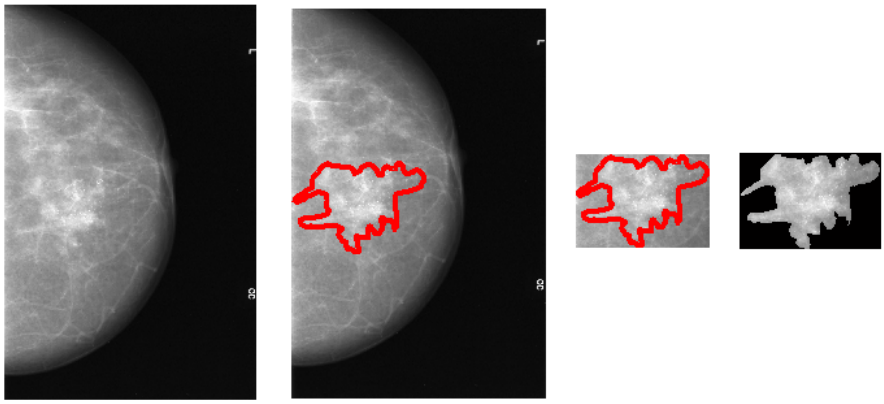


Figure 5.1: The images obtained during the pre-processing. From the left to right there are the following images: First the image of the whole breast, then the image with the polygon surrounding the finding, then the cropped image of the finding and at last the finding without any background pixel.

- **Age**: of the patient during the exam.
- **Breast density**: of the patient according to the BI-RADS standard.
- **Mammography Nodule**: a boolean value if there is a mass.
- **Mammography Calcification**: a boolean value if calcifications were detected.
- **Mammography Microcalcification**: a boolean value if microcalcifications were detected.
- **Mammography Axillary Adenopathy**: a boolean value if axillary adenopathy was detected.
- **Mammography Architectural Distortion**: a boolean value if there are signs of architectural distortion.
- **Mammography Stroma Distortion** a boolean value if there are signs of stroma distortion.

5.2.2 Intensity

This data was obtained directly from the dataset, the BCDR-F01 had a text file with intensity descriptors calculated of the findings of each image. This subset will be mentioned as *S2* and contains the following features, where n is the number of pixels of the finding and x_i intensity value of the i^{th} pixel of the finding:

- **Mean**: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- **Standard Deviation**: $\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- **Skewness**: $\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^3}$
- **Kurtosis**: $\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2} - 3$
- **Minimum**: is the minimum x_i value of the finding.
- **Maximum**: is the maximum x_i value of the finding.

5.2.3 Wavelets

Wavelets are also common descriptors used in breast cancer classification, an example of a wavelet descriptor (Gabor) was discussed at Subsection 2.1, we also extracted wavelets from the BCDR dataset using the matlab toolbox of signal processing with the default parameters, such as Haar wavelets with 4 scales of resolution, to analyse them in our study. We did not explore other wavelet functions or wavelet parameters because there are several functions and parameters, and focusing on which is the best image descriptor with the best parametrization is not our goal, we aim to study the overall performance of MKL for breast cancer data. And for that purpose, extracting one wavelet suffices, we will refer this wavelet extracted as *S3*.

5.2.4 Local Binary Patterns (LBP)

While analysing the literature on breast cancer classification using medical images, we noticed one work with great results using LBP for this purpose [56]. Therefore, we decided to also test them in our experimental study. In order to extract the LBP from the database used, we used the vlFeat Library [74], which calculates the LBP descriptor as it follows: First the image is separated into cells (e.g. if cell size equals 2 then the image is separated into cells with half width and half height) then for each cell it is calculated a string of bits for each pixel with a 3x3 neighbourhood where each bit is turned on if the intensity value of the neighbour pixel is higher than the

intensity value of the central pixel or turned off otherwise. Starting from the pixel in the right of the central pixel and following with the next pixel according to a clockwise direction. Since a string of 8 bits may result in a variation of 256 possible patterns, this library uses an uniform quantization method to reduce the number of LBPS to 58 quantized patterns as illustrated in Figure 5.2. With the calculated quantized patterns, a frequency histogram is created and normalized. Resulting in a normalized histogram of quantized patterns for each cell. These histograms are aggregated into one normalized histogram that will be used as the feature vector, containing 58 floats. In order to study the LBP with breast cancer data, we generated 5 subsets of local binary patterns with different cellsizes. Depending on the cell size m, each dataset will be addressed as it follows: $m = 2$ as *S4*, $m = 4$ as *S5*, $m = 6$ as *S6*, $m = 8$ as *S7* and $m = 16$ as *S8*. There were some cases, specially for the higher cell size, where the region of interest extracted would not contain enough pixels to divide in regions according to the cell size, for these cases we did a 2 times upscale on the region of interest before applying the LBP.

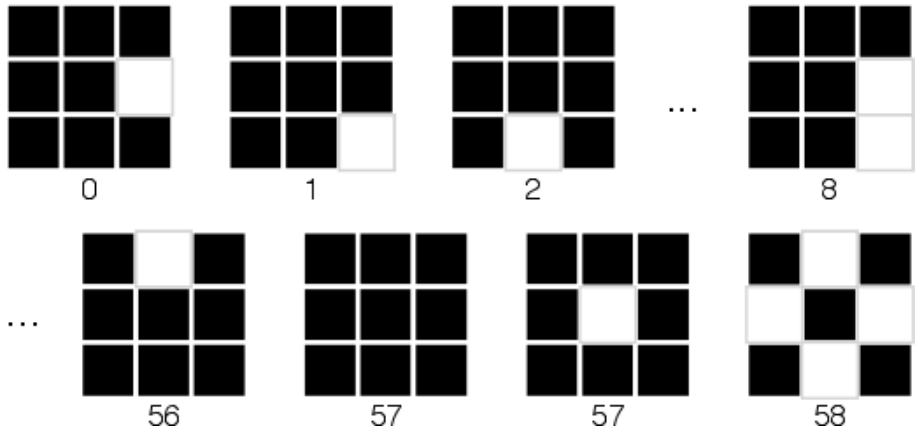


Figure 5.2: The 58 uniform quantized local binary patterns used to build the LBP quantized histogram.

5.2.5 Histogram of Gradient Divergence (HGD)

During the literature review, this descriptor obtained one of the best results regarding breast cancer classification. It was developed by [4], to extract information regarding breast cancer images, specially for masses. The code (in Matlab) to extract the HGD was kindly provided by the authors. A brief example of how this descriptor works is illustrated in Figure 5.3, for in depth information consult [4]. In order to extract the histogram of gradient divergence we used the following parameters, the histogram

normalization tnorm = 2, the number of bins for direction nbins = 8 and selecting 3 as the number of regions on the HGD (nx) for each direction x. Thus, obtaining a feature vector of the normalized sum of the histograms of size 24 floats (nbins × nx). We will address this subset of patterns as *S9*.

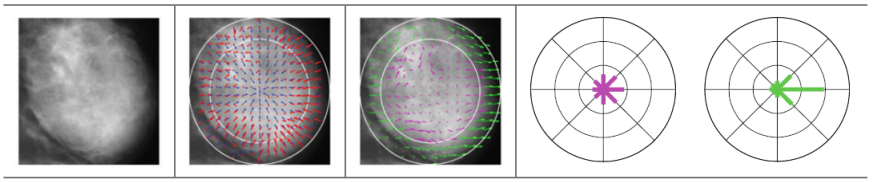


Figure 5.3: The histogram of gradient divergence from a mass with well-defined borders. On the first image we have raw image of the mass. On the second image we have a sparse representation of the gradient (red arrows) and the reference (convergence) vectors (blue arrows), that will be used to calculate the gradient divergence vectors represented on the third image, which have magnitude equal to the gradient and orientation equal to the angular difference between the gradient and the reference vector (horizontal, left to right vectors means zero divergence). Then, for each region (the center region and the border region) it is calculated a 8 direction bins (where zero divergence points to the right, and the remaining following anti-clockwise) resulting in a histogram for each zone as we can see on the last image. The decriptor is represented in a vector of 16 (8+8) values of each bin. This image was taken from [4]

5.3 Classification model

For this experiment, we will use two classification models. The MKL and the SVM. As we explained on the previous chapter, both are kernel methods. In order to compare them we perform the experiment selecting the same kernel for both methods, a Polynomial Kernel of degree 1 which is the same as a Linear Kernel. We also explored further the MKL trying the usage of a kernel with proven performance on image classification [75], in order to study if it would perform well with the breast cancer image data extracted.

- **Polynomial Kernel** (Linear): This kernel is represented by the formula 5.1, receiving as parameter the degree. In this experiment we wanted to compare the linear kernel, i.e. degree (d) equals 1.
- **Heavy-Tailed Radial Basis Kernel** (HTRBF): This kernel is represented by the formula 5.2, receiving as parameters the boundaries a and b. Since the MKL

allows to the user to add all the parameter values they want to test, selecting the most appropriated parameters to generate the final model, we used as parameter all the combinations with a = 0.1 or 0.5 or 1.0 and b = 0.2 or 1.0 or 2.0, i.e. the following set of parameters [a , b]: [[0.1, 0.2];[0.1, 1.0];[0.1, 2.0];[0.5, 0.2];[0.5, 1.0];[0.5, 2.0];[1.0, 0.2];[1.0, 1.0];[1.0, 2.0]].

$$K(x, y) = (x.y)^d \quad \text{where d} = 1 \tag{5.1}$$

$$K(x, y) = exp(-\frac{\|x^a - y^a\|^b}{2\sigma^2}) \tag{5.2}$$

Resulting in three classification models for the experimental study that will be addressed as it follows: The Linear SVM as *C1*, the Linear MKL as *C2* and the Heavy-Tailed RBF MKL as *C3*.

5.4 Experimental Study

In order to evaluate the features extracted *S1*, *S2*, ..., *S9* on each view (CC and MLO) and the classification models that we selected to compare *C1*, *C2* and *C3*, we decided to perform independently for each classifier and feature set, 50 repetitions of the Holdout method illustrated in Figure 5.4 for each view. Using resampling without replacement on each repetition to split 80% of the data to be used as train set and 20% to be used as test set. In order to select the most appropriate parameter for each classifier method (all three methods require the constant C parameter, explained on the previous Chapter) on each repetition, we performed an unstratified three fold cross validation using only the train set for each of the C parameters ranging from 10^{-2} to 10^3, selecting the smaller C with highest AUC to be used to train the whole train set. For each repetition we saved the table containing the true values from the respective test set and the respective predicted values. We also used this table to calculate the AUC, and used as final metric to analyse each combination of classifier and feature set, the average of the AUCs calculated, and the standard deviation of these AUCs. Since these results are for each view, we also did the average of the results of each view to obtain an overall result for both views. After performing the experiment for each set, we selected the two feature sets with the highest performance (*S9* and *S4*) to explore if they can perform better with the concatenation of clinical data, these new sets will be referred as *S1+S9* and *S1+S4*, and to also explore how they perform with only masses, i.e. a subsets extracted from *S9* and *S4* by removing any row that represented a finding that was not a mass, these subsets will be referred as $S9_m$ and

$S4_m$. The same experiment described above was performed for these new feature sets.

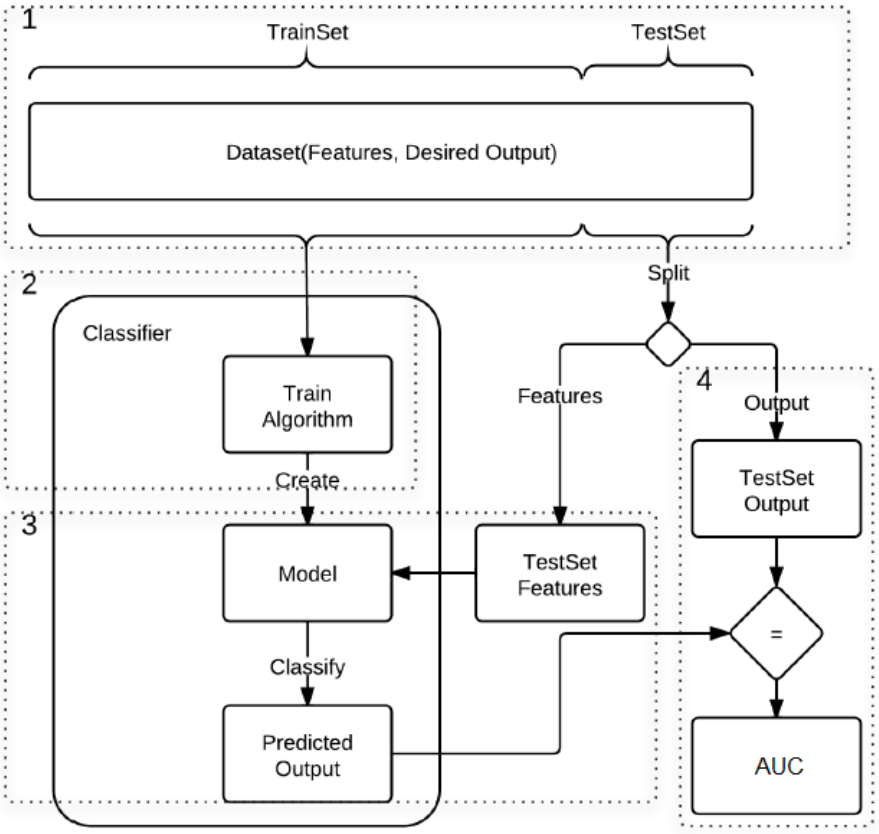


Figure 5.4: The Holdout method to evaluate a classifier. First the dataset is divided in 2 subsets, one for training, other for tests. We use the train dataset to create an classification model. After we send the test set features to the model which will return the classification answers and compare those predicted answers with the real answers from the test set. Allowing to calculate the metrics such as AUC.

5.5 Results and Discussion

In order to obtain an overview of all the results, we created a scatter plot 5.5, containing the AUC and AUC Standard Deviation (SD) of each experiment. From this scatter plot we can see clearly that the Simple MKL surpass the SVM in almost all types of image patterns on breast cancer images. We can also notice that the Local Binary Patterns and Histogram of Gradient Divergence, obtained the best results. In order to see in more detail the results, we listed in 5 tables all the results in AUC obtained separated by each view since the experiment was performed independently on each view, using the average of the AUC of both views as the final result. Where

on the first 3 tables, Table 5.1, Table 5.2 and Table 5.3 we list the results obtained for each view for each of the three classifiers *C1*, *C2* and *C3*, where the best results were obtained by *C2*, reaching an AUC of 0.82. Then the results of the experiment using only the image descriptors with better AUC, are displayed in Table 5.4 and Table 5.5 where on the first table we analyse them with the addition of clinical data and on the second table we show the results using only masses as data. In summary, MKL clearly outperforms the SVM with the same kernel. Although we would like to compare with other studies from the literature, most of the works are not comparable due the difference on the experiment procedure done or on the datasets used. Nevertheless, we based our experiment procedure on Moura's work [4] which allows us to compare our results obtained. Despite the fact that he used SVM from SMO while we used SVM from LIBSVM which might be the reason for our different results with the Linear SVM, we obtained better results with the Simple MKL method using the HGD in two experiments, on the HGD for masses only and on the HGD for all findings combined with clinical data.

MKL HTRBF	CC View	MLO View	CC MLO Average
Clinical	0.707	0.668	0.688
Intensity	0.587	0.622	0.605
Wavelet	0.614	0.627	0.621
LBP (m = 2)	0.787	0.735	0.787
LBP (m = 4)	0.764	0.764	0.764
LBP (m = 6)	0.789	0.759	0.774
LBP (m = 8)	0.784	0.751	0.768
LBP (m = 16)	0.809	0.744	0.776
HGD	0.783	0.827	0.805

Table 5.1: This table shows the AUC results of the experimental study using MKL with the HTRBF kernel for all subsets of image features for each of the views for all findings (CC and MLO)

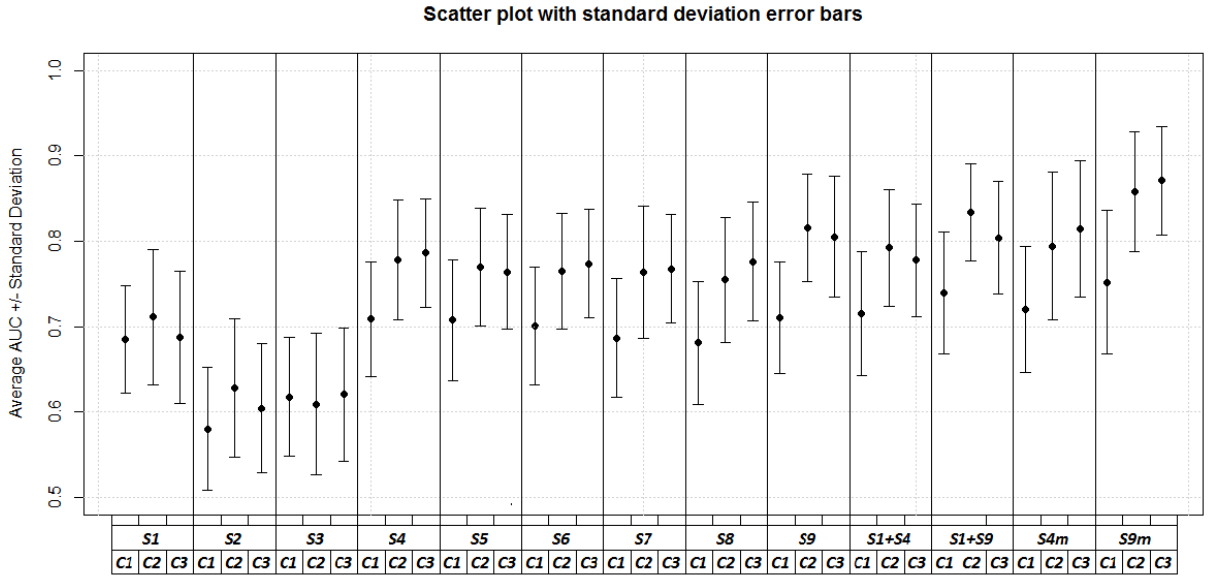


Figure 5.5: This scatter plot displays the results of the experiment, AUC (points) and the AUC standard deviation (bars), using each classifier method and each image pattern descriptor. Scatter plot legend: *S1* Clinical Data, *S2* Intensity, *S3* Wavelets, *S4* Local Binary Pattern with $m = 2$, *S5* Local Binary Pattern with $m = 4$, *S6* Local Binary Pattern with $m = 6$, *S7* Local Binary Pattern with $m = 8$, *S8* Local Binary Pattern with $m = 16$, *S9* Histogram of Gradient Divergence, $S4_m$ Local Binary Pattern with mass only and $m = 2$, $S9_m$ Histogram of Gradient Divergence with mass only, *S1+S4* Clinical Data plus Local Binary Pattern with $m = 2$, *S1+S9* Clinical Data plus Histogram of Gradient Divergence, *C1* Multiple Kernel Learning classifier with Heavy-Tailed RBF Kernel, *C2* Multiple Kernel Learning classifier with Linear Kernel, *C3* Support Vector Machine with Linear Kernel.

MKL Linear	CC View	MLO View	CC MLO Average
Clinical	0.717	0.706	0.711
Intensity	0.589	0.668	0.628
Wavelet	0.617	0.602	0.609
LBP (m = 2)	0.793	0.762	0.778
LBP (m = 4)	0.784	0.755	0.770
LBP (m = 6)	0.800	0.730	0.765
LBP (m = 8)	0.788	0.740	0.764
LBP (m = 16)	0.798	0.712	0.755
HGD	0.820	0.813	0.816

Table 5.2: This table shows the AUC results of the experimental study using MKL with the Linear kernel for all subsets of image features for each of the views for all findings (CC and MLO)

SVM Linear	CC View	MLO View	CC MLO Average
Clinical	0.709	0.661	0.685
Intensity	0.544	0.622	0.583
Wavelet	0.646	0.586	0.616
LBP (m = 2)	0.716	0.702	0.709
LBP (m = 4)	0.716	0.699	0.708
LBP (m = 6)	0.708	0.694	0.701
LBP (m = 8)	0.694	0.679	0.687
LBP (m = 16)	0.702	0.660	0.681
HGD	0.721	0.700	0.710

Table 5.3: This table shows the AUC results of the experimental study using SVM with the Linear kernel for all subsets of image features for each of the views for all findings (CC and MLO)

Classification Method	SVM	MKL	MKL
Kernel Selected	Linear	Linear	HTRBF
LBP m2 + Clinical	0.715	0.792	0.778
HGD + Clinical	0.739	0.834	0.804

Table 5.4: Comparison between SVM linear, MKL Linear and MKL HTRBF using feature sets such as LBP (m = 2) and HGD combined with clinical data

Classification Method	SVM	MKL	MKL
Kernel Selected	Linear	Linear	HTRBF
LBP m2	0.721	0.794	0.815
HGD	0.752	0.858	0.871

Table 5.5: This table compares the results between SVM linear, MKL Linear and MKL HTRBF using feature sets such as LBP (m = 2) and HGD extracted only from masses.

Chapter 6

Conclusions and Future Work

Each year that pass we are getting closer to a better system for breast cancer detection and classification. Each improvement may be little but might save many lives. We know that the current systems are efficient on classification, although they are not perfect. We believe that with our experimental study, we showed that MKL may improve the results in comparison with the most used method on breast cancer anomaly classification, the SVM. We hope to have convinced the reader with the work we presented in this document.

6.1 Research summary

In this work, we have studied in depth the current systems used for breast cancer classification that are mainly composed by three parts, the breast cancer databases, the image processing methods for pattern extraction and the machine learning methods used to classify the patterns extracted. In the database area, we described all datasets are available (as far as we know), listing their characteristics and explaining which characteristics are considered optimal, and we selected one database to perform our experiment. Then on the area of processing breast cancer images, we studied each of the methods used, and we integrated within our experiment the methods with good performance in the literature that had currently available libraries. Then, on the classification area, we searched for the most used method, and also searched for several types of approaches. Since MKL was not introduced in this area, we did a brief explanation of a simple version of MKL, the Simple MKL and performed an experimental study to introduce it in this area and compare it with the most used classifier

method, using patterns extracted from the selected dataset (patterns given by the previously integrated image processing methods). In summary, with this dissertation we explored the methods in the literature for breast cancer classification, introduced and explained the Simple MKL method and performed a comparison study between SVM and Simple MKL using several image patterns, obtaining statistical meaningful results that may be used as base to motivate further studies in this area with more complex and efficient MKL approaches. Therefore, we believe we accomplished the goal of presenting a method that can improve the current performance of breast cancer classification by introducing and analysing the usage of MKL in the breast cancer classification field.

6.2 Main findings

We can show that from the analysed state of the art (Chapter 3) automatic breast cancer classification is a hot topic due to the diversity of image pattern extraction and machine learning methods used. We can also show that the works with higher performance used one of the following image pattern extraction methods, Wavelets, Histogram of Gradient Divergence and Local Binary Patterns. Where most of the times they recurred to simple methods such as kNN or SVM to classify the patterns extracted leaving space for improvement although there are some exceptions that we also listed. However, MKL is an interesting classification method given its capacity on fitting data with more flexibility than the SVM method as we explained in Chapter 4. Therefore, we performed an experiment using MKL and found that the method was indeed a way to improve the performance on this area. Obtaining better performance in the classification of all lesions in comparison with the Linear SVM with most of the tested breast cancer patterns extracted, where on the best result using Histogram of Gradients, the MKL method obtained the AUC of 0.82 in comparison with the AUC of 0.71 obtained from the SVM. We also found that the usage of clinical data combined with image patterns descriptors enhance the overall performance, during our experiment we obtained an increment of 0.02 the AUC of our best result. And our last main finding was that the lesion type mass contain many cues regarding malignancy, given the great results obtained on our experiment using only masses, achieving an AUC of 0.87.

6.3 Current limitations

In order to perform our comparison experiment, we used only the BCDR F01 dataset, despite being a good dataset as we discussed in Chapter 3, the dataset does not contain many rows for each lesion, we were only able to analyse the classification of all lesions and the classification of masses. The dataset also contains a good number of micro calcifications but our method of region of interest pattern extraction does not explore details that are relevant in micro calcification classification, such as the number of micro calcifications, their distance, their position and others, therefore we did not attempt to use them separately on our experiment like we did with masses. Another limitation of using only this dataset was the fact that we could only compare our results with works that used it, which it is not a real issue since our goal was to prove the that MKL is a good approach as a classifier for breast cancer data and not to compare results with the methods in the state of the art.

6.4 Future work

Now that we performed the initial step to the usage of MKL in the breast cancer classification, we hope to motivate new researchers to explore further in three directions. The first direction would be to study the top performance MKL methods in the literature to build a mass classification model, because we believe that great results that may surpass the state of the art results in breast cancer mass classification can be achieved if the HGD descriptor (which has exceptional performance in masses) is used to study which MKL classifier method has the best performance. We only used one dataset in our study, the second direction would be to replicate the same experiment for other datasets and see if there are variations on the comparison results. Since the MKL has a bigger adaptability to data (because it's based on the creation of a optimal kernel to fit that data), merging two different datasets may not compromise the results, which would lead to a classifier able to classify between images taken from different scanners which is great for real case scenarios, because many hospitals have different equipments. Last, we believe that a separated study of different types of lesions may be relevant, although may be hard to perform because some lesions are more rare than others, therefore its hard to find datasets with plenty cases of each type of finding to perform a good study, but exploring these may lead to the best results in classification of breast cancer images.

6.5 Conclusion

The proposed method, MKL can get a noticeable performance in breast cancer image data, and should be further explored since it is a more flexible approach to classify data than the most common method used in the literature, the SVM. Even though our experiment, which used only a simple method of MKL, the Simple MKL, part of our results surpassed some of the results listed in the literature. And we also achieved our goal, since our experiment proved the MKL superiority over the SVM for breast cancer data. Also encouraging new works to pursue more complex MKL methods to obtain results that may improve the current state of the art in breast cancer classification.

References

[1] Daniel C Moura and Miguel A Guevara López. An evaluation of image descriptors combined with clinical data for breast cancer diagnosis. *International journal of computer assisted radiology and surgery*, pages 1–14, 2013. xvii, 2, 3, 8, 30, 32, 33, 34, 35, 36

[2] Mammogram image, May 2014. xx, 20

[3] Ellen Shaw De Paredes. *Atlas of mammography*. Lippincott Williams & Wilkins, 2007. xx, 22

[4] Daniel C Moura and Miguel A Guevara López. An evaluation of image descriptors combined with clinical data for breast cancer diagnosis. *International journal of computer assisted radiology and surgery*, 8(4):561–574, 2013. xxi, 51, 52, 55

[5] World Health Organization International Agency for Research on Cancer. Latest world cancer statistics, press release: N 223 dec. 12, 2013. 1

[6] Marina Velikova, Inês Dutra, and Elizabeth S Burnside. Automated diagnosis of breast cancer on medical images, 2013. 1

[7] Fiona J Gilbert, Susan M Astley, Magnus A McGee, Maureen GC Gillan, Caroline RM Boggis, Pamela M Griffiths, and Stephen W Duffy. Single reading with computer-aided detection and double reading of screening mammograms in the united kingdom national breast screening program1. *Radiology*, 241(1):47–53, 2006. 1

[8] Alain Rakotomamonjy, Francis Bach, Stéphane Canu, Yves Grandvalet, et al. Simplemkl. *Journal of Machine Learning Research*, 9:2491–2521, 2008. 1, 3, 39, 43, 44, 45

[9] Mohamed Meselhy Eltoukhy, Ibrahima Faye, and Brahim Belhaouari Samir. A comparison of wavelet and curvelet for breast cancer diagnosis in digital

mammogram. *Computers in Biology and Medicine*, 40(4):384–391, 2010. 2, 32, 33, 34, 35

[10] Gensheng Zhang, Wei Wang, Jucheol Moon, Jeong K Pack, and Soon Ik Jeon. A review of breast tissue classification in mammograms. In *Proceedings of the 2011 ACM Symposium on Research in Applied Computation*, pages 232–237. ACM, 2011. 2, 32, 34

[11] Mary M Galloway. Texture analysis using gray level run lengths. *Computer graphics and image processing*, 4(2):172–179, 1975. 8

[12] Aswini Kumar Mohanty, Pratap Kumar Champati, Sukanta Kumar Swain, and Saroj Kumar Lenka. A review on computer aided mammography for breast cancer diagnosis and classification using image mining methodology. *International Journal of Computer Science and Communication*, 2(2):531–538, 2011. 8, 32

[13] Rafael C Gonzalez and Richard E Woods. Digital image processing, 2002. 9

[14] David Bavrina. Gabor wavelets in image processing. In *Proceedings of the 17th Conference STUDENT EEICT 2011*, pages 522–526. Brno University of Technology, 2011. 9

[15] Tinne Tuytelaars and Krystian Mikolajczyk. Local invariant feature detectors: a survey. *Foundations and Trends® in Computer Graphics and Vision*, 3(3):177–280, 2008. 9

[16] David G Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999. 10, 33

[17] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27, 1967. 11

[18] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984. 11, 13

[19] Jason D Rennie, Lawrence Shih, Jaime Teevan, David R Karger, et al. Tackling the poor assumptions of naive bayes text classifiers. In *ICML*, volume 3, pages 616–623. Washington DC), 2003. 11, 15

[20] Christopher M Bishop et al. *Pattern recognition and machine learning*, volume 1. springer New York, 2006. 11, 40, 41

[21] Tan Pang-Ning, Michael Steinbach, Vipin Kumar, et al. Introduction to data mining. *WP Co*, 2006. 13

[22] Laurent Hyafil and Ronald L Rivest. Constructing optimal binary decision trees is np-complete. *Information Processing Letters*, 5(1):15–17, 1976. 13

[23] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996. 13

[24] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. 13

[25] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. 13, 41

[26] Irina Rish. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46, 2001. 15

[27] Charles E Metz. Basic principles of roc analysis. In *Seminars in nuclear medicine*, volume 8, pages 283–298. Elsevier, 1978. 16

[28] Erin LeDell, Maya L Petersen, and Mark J van der Laan. Computationally efficient confidence intervals for cross-validated area under the roc curve estimates. 2012. 17

[29] Juan D Rodriguez, Aritz Perez, and Jose A Lozano. Sensitivity analysis of k-fold cross validation in prediction error estimation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(3):569–575, 2010. 17

[30] George Forman and Martin Scholz. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *ACM SIGKDD Explorations Newsletter*, 12(1):49–57, 2010. 18

[31] Suryanarayan S. Karellas A, Vedantham S. Digital mammography: an emerging technology. *Business Briefing: Future Directions in Imaging*, 2006. 19

[32] Etta D Pisano, Constantine Gatsonis, Edward Hendrick, Martin Yaffe, Janet K Baum, Suddhasatta Acharyya, Emily F Conant, Laurie L Fajardo, Lawrence Bassett, Carl D'Orsi, et al. Diagnostic performance of digital versus film mammography for breast-cancer screening. *New England Journal of Medicine*, 353(17):1773–1783, 2005. 19, 30

[33] Judy Chavez RT(R)(M)(BS) RDMS Muriel Anita Simmons, RT(R)(M) and DO Lora D. Barke. Patient positioning and mammography. 2012. 19

[34] William E Barlow, Constance D Lehman, Yingye Zheng, Rachel Ballard-Barbash, Bonnie C Yankaskas, Gary R Cutter, Patricia A Carney, Berta M Geller, Robert Rosenberg, Karla Kerlikowske, et al. Performance of diagnostic mammography for women with signs or symptoms of breast cancer. *Journal of the National Cancer Institute*, 94(15):1151–1159, 2002. 19

[35] Daniel B Kopans. *Breast imaging*. Lippincott Williams & Wilkins, 2007. 21

[36] C. J. D. Orsi, L. W. Bassett, W. A. Berg, and et al. Bi-rads: Mammography, 4th edition. In *4th edition*. American College of Radiology, Inc., 2003. 23

[37] Arnau Oliver, Jordi Freixenet, Joan Martí, Elsa Pérez, Josep Pont, Erika RE Denton, and Reyer Zwiggelaar. A review of automatic mass detection and segmentation in mammographic images. *Medical Image Analysis*, 14(2):87–110, 2010. 25

[38] John Suckling, J Parker, DR Dance, S Astley, I Hutt, C Boggis, I Ricketts, E Stamatakis, N Cerneaz, Siew-Li Kok, et al. The mammographic image analysis society digital mammogram database. 1994. 27, 32

[39] K Bowyer, D Kopans, WP Kegelmeyer, R Moore, M Sallam, K Chang, and K Woods. The digital database for screening mammography. In *Third International Workshop on Digital Mammography*, volume 58, 1996. 27, 28

[40] Michael Heath, Kevin Bowyer, Daniel Kopans, P Kegelmeyer Jr, Richard Moore, Kyong Chang, and S Munishkumaran. Current status of the digital database for screening mammography. In *Digital mammography*, pages 457–460. Springer, 1998. 27, 28

[41] Júlia EE Oliveira, Mark O Gueld, Arnaldo de A Araújo, Bastian Ott, and Thomas M Deserno. Toward a standard reference database for computer-aided mammography. In *Medical Imaging*, pages 69151Y–69151Y. International Society for Optics and Photonics, 2008. 27, 29, 32

[42] Thomas M Deserno, Michael Soiron, Júlia EE de Oliveira, and Arnaldo de A Araújo. Computer-aided diagnostics of screening mammography using content-based image retrieval. In *SPIE Medical Imaging*, pages 831527–831527. International Society for Optics and Photonics, 2012. 27, 29

[43] Inês C Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria João Cardoso, and Jaime S Cardoso. Inbreast: toward a full-field digital mammographic database. *Academic radiology*, 19(2):236–248, 2012. 27, 30, 31, 32

[44] Ineji. Breast cancer digital repository, May 2014. 27, 30, 32

[45] LAPIMO. Bancoweb lapimo database repository, May 2014. 27, 29, 32

[46] RW Blamey, B Hornmark-Stenstam, G Ball, M Blichert-Toft, L Cataliotti, A Fourquet, J Gee, K Holli, R Jakesz, M Kerin, et al. Corrigendum to oncopool–a european database for 16,944 cases of breast cancer[european journal of cancer 46 (2009) 56–71]. *European Journal of Cancer*, 46(9):1762, 2010. 27, 31

[47] Sameer Singh and Keir Bovis. An evaluation of contrast enhancement techniques for mammographic breast masses. *Information Technology in Biomedicine, IEEE Transactions on*, 9(1):109–119, 2005. 28

[48] Mehul P Sampat, Alan C Bovik, Gary J Whitman, and Mia K Markey. A model-based framework for the detection of spiculated masses on mammographya). *Medical physics*, 35(5):2110–2123, 2008. 28

[49] L. L. N. Laboratories and University of California S.F. Radiology Department. Llnl ucsf, May 2014. 31, 32

[50] University of South Florida. Digital mammography other resources, May 2014. 31

[51] I Domingues, E Sales, JS Cardoso, WCA Pereira, and Programa de Engenharia Biomédica-COPPE. Inbreast-database masses characterization. 32

[52] Sung-Nien Yu, Kuan-Yuei Li, and Yu-Kun Huang. Detection of microcalcifications in digital mammograms using wavelet filter and markov random field model. *Computerized Medical Imaging and Graphics*, 30(3):163–173, 2006. 32, 33

[53] E Sakka, A Prentza, IE Lamprinos, and D Koutsouris. Microcalcification detection using multiresolution analysis based on wavelet transform. In *IEEE International Special Topic Conference on Information Technology in Biomedicine, Ioannina, Epirus, Greece*, 2006. 32, 33

[54] Hamid Soltanian-Zadeh, Farshid Rafiee-Rad, and Siamak Pourabdollah-Nejad D. Comparison of multiwavelet, wavelet, haralick, and shape features for microcalcification classification in mammograms. *Pattern Recognition*, 37(10):1973–1986, 2004. 32, 33

[55] Subodh Srivastava, Neeraj Sharma, Sanjay Kumar Singh, and Rajeev Srivastava. Design, analysis and classifier evaluation for a cad tool for breast cancer detection

from digital mammograms. *International Journal of Biomedical Engineering and Technology*, 13(3):270–300, 2013. 33

[56] Mohamed A Berbar, Yaser A Reyad, and Mohamed Hussain. Breast mass classification using statistical and local binary pattern features. In *Information Visualisation (IV), 2012 16th International Conference on*, pages 486–490. IEEE, 2012. 33, 50

[57] Defeng Wang, Lin Shi, and Pheng Ann Heng. Automatic detection of breast cancers in mammograms using structured support vector machines. *Neurocomputing*, 72(13):3296–3302, 2009. 33, 34, 36

[58] Sumeet Dua, Harpreet Singh, and Hilary W Thompson. Associative classification of mammograms using weighted rules. *Expert systems with applications*, 36(5):9250–9259, 2009. 33, 34, 35, 37

[59] Essam A Rashed, Ismail A Ismail, and Sherif I Zaki. Multiresolution mammogram analysis in multilevel decomposition. *Pattern Recognition Letters*, 28(2):286–292, 2007. 33

[60] Rafayah Mousa, Qutaishat Munib, and Abdallah Moussa. Breast cancer diagnosis system based on wavelet analysis and fuzzy-neural. *Expert systems with Applications*, 28(4):713–723, 2005. 33, 35, 37

[61] Cristiane Bastos Rocha Ferreira and Dıbio Leandro Borges. Analysis of mammogram classification using a wavelet transform decomposition. *Pattern Recognition Letters*, 24(7):973–982, 2003. 33, 34, 35

[62] Ioan Buciu and Alexandru Gacsadi. Directional features for automatic tumor classification of mammogram images. *Biomedical Signal Processing and Control*, 6(4):370–378, 2011. 33, 34, 36

[63] Rodrigo Pereira Ramos, Marcelo Zanchetta do Nascimento, and Danilo Cesar Pereira. Texture extraction: An evaluation of ridgelet, wavelet and co-occurrence based methods applied to mammograms. *Expert Systems with Applications*, 39(12):11036–11047, 2012. 33, 35, 36

[64] Mohamed Meselhy M Eltoukhy, I Faye, and Brahim Belhaouari Samir. Using curvelet transform to detect breast cancer in digital mammogram. In *Signal Processing & Its Applications, 2009. CSPA 2009. 5th International Colloquium on*, pages 340–345. IEEE, 2009. 33, 34

[65] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005. 33

[66] VD Nguyen, DT Nguyen, TD Nguyen, QD Truong, and MD Le. Combination of block difference inverse probability features and support vector machine to reduce false positives in computer-aided detection for massive lesions in mammographic images. In *Biomedical Engineering and Informatics (BMEI), 2013 6th International Conference on*, pages 28–32. IEEE, 2013. 34, 36

[67] Nabiha Azizi, Yamina Tlili-Guiassa, and Nawel Zemmal. A computer-aided diagnosis system for breast cancer combining features complementarily and new scheme of svm classifiers fusion. *International Journal of Multimedia & Ubiquitous Engineering*, 8(4), 2013. 34, 36, 37

[68] Emina Alickovic and Abdulhamit Subasi. Comparison of decision tree methods for breast cancer diagnosis. 2013. 35, 36

[69] Mahjabeen Mirza Beg and Monika Jain. An analysis of the methods employed for breast cancer diagnosis. *arXiv preprint arXiv:1206.3777*, 2012. 35, 36

[70] Zhimin Huo, Maryellen L Giger, Carl J Vyborny, Dulcy E Wolverton, Robert A Schmidt, and Kunio Doi. Automated computerized classification of malignant and benign masses on digitized mammograms. *Academic Radiology*, 5(3):155–168, 1998. 35, 37

[71] A Aizerman, Emmanuel M Braverman, and LI Rozoner. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and remote control*, 25:821–837, 1964. 40

[72] N. Aronszajn. Theory of reproducing kernels. In *Trans. Am. Math.*, volume 68, pages 337–404. Soc., 1950. 41, 42

[73] David G Luenberger and Yinyu Ye. *Linear and nonlinear programming*, volume 116. Springer, 2008. 44

[74] Vlfeat library, 2014. 50

[75] Olivier Chapelle, Patrick Haffner, and Vladimir N Vapnik. Support vector machines for histogram-based image classification. *Neural Networks, IEEE Transactions on*, 10(5):1055–1064, 1999. 52