# Weighted Multiple Kernel Learning for Breast Cancer Diagnosis applied to Mammograms


Tiago André Guedes Santos

Mestrado Integrado em Engenharia de Redes e Sistemas Informáticos

Departamento de Ciência de Computadores

2016/2017

**Orientador**

Inês de Castro Dutra, Professora Auxiliar,

Faculdade de Ciências da Universidade do Porto

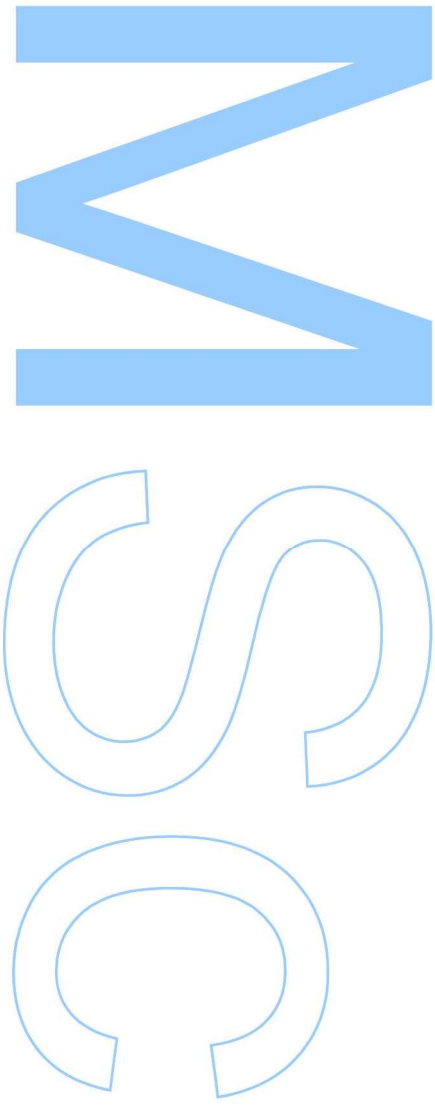
U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

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

O Presidente do Júri,

Porto, _____/_____/_____



# Abstract

Detecting breast cancer in mammograms can be a hard task even to most experienced specialists. Several works in the literature have tried to build models to describe malignant or benign findings using BI-RADS annotated features or features automatically extracted from images. Some of the best models are based on Support Vector Machines (SVMs). Features from mammograms have heterogeneous types and most methods handle them equally. Multiple Kernel Learning (MKL) can create models where each feature can be treated in a different way, which may improve the quality of the learned models. In this work, we use MKL to help building models to distinguish between malignant and benign findings. One of the problems with this domain is that the classes are unbalanced: fortunately the number of malignant cases is much smaller than the number of benign cases. However, this imbalance may lead an MKL classifier to label most of the cases as benign. We improve on these models by adopting a strategy of weighing the benign and malignant cases in order to produce models that are more reliable and robust to the class distribution. Our results show that our weighted approach produces better quality models for both balanced and unbalanced mammogram datasets.

# Resumo

Detetar casos de cancro da mama em mamografias pode ser uma tarefa difícil mesmo para os especialistas mais experientes. Vários trabalhos na literatura têm vindo a construir modelos para classificar regiões de interesse, usando atributos provenientes de anotações BI-RADS ou extraídos automaticamente de imagens. Alguns dos melhores modelos são baseados em Support Vector Machines (SVMs). Os atributos utilizados neste tipo de problema têm sempre tipos heterogéneos, e a maioria dos métodos trata todos estes da mesma maneira. O Multiple Kernel Learning (MKL) pode criar modelos onde cada tipo de atributo pode ser tratado de forma diferente, o que pode melhorar a qualidade dos modelos aprendidos. Neste trabalho, usamos o MKL para construir modelos que classificam regiões de interesse como benignas e malignas. Um dos problemas com este domínio é que as classes são desequilibradas: felizmente, o número de casos malignos é muito menor que o número de casos benignos. No entanto, esse desequilíbrio pode levar um classificador MKL a classificar a maioria dos casos como benignos. Melhoramos estes modelos, adotando uma estratégia que aplica peso aos casos benignos e malignos, para assim produzir modelos mais confiáveis e robustos na distribuição das classes. Os resultados mostram que nossa abordagem produz modelos de melhor qualidade para datasets balanceados ou não balanceados.

# Contents

# List of Tables

# List of Figures

# List of Algorithms

# Acronyms

**AUC** Area Under the Curve

**ANN** Artificial Neural Networks

**ANOVA** Analysis Of Variance

**BCDR** Breast Cancer Digital Repository

**BI-RADS** Breast Imaging-Reporting and Data System

**BN** Bayesian-Network

**CAD** Computer-Aided diagnosis

**CSSVM** Cost-Sensitive Support Vector Machine

**DDSM** Digital Database for Screening Mammography

**FOR** False Omission Rate

**GLCM** Gray-Level Co-occurrence Matrix

**HTRBF** Heavy Tailed Radial Basis Function

**KRR** Kernel Ridge Regression

**MKL** Multiple Kernel Learner

**NLMKL** Non-Linear Multiple Kernel Learner

**PR** Precision

**RBF** Radial Basis Function

**REC** Recall

**ROC Curve** Receiver Operating Characteristic Curve

**ROI** Region Of Interest

**SVM** Support Vector Machine

# Chapter 1

# Introduction

## 1.1 Motivation

Breast cancer is one of the most common forms of cancer. A mammogram, or X-ray of the breast, is a popular technique used to detect cancer at an early stage. If some suspicious finding is found in a mammogram, a biopsy is usually recommended in order to decide on surgery. Biopsy is a necessary, but also aggressive, high-stakes procedure. Although statistics vary among publications and countries, depending on age and other conditions, overall, according to the National Institutes of Health (NIH) a specialist misses 20% of breast cancers that are present at the time of screening. Usually, the main reason for that is high breast density. Some of these missing cancers can be detected through clinical physical exam of the breast. According to Hofvind et al. (2012) the false positive rate ranges from 8% to 21%, depending on the patient's age. Patients with a false negative result will have a false sense of security and potential delay in cancer diagnosis, while patients with a false positive result will go through additional testing (very often, intrusive and aggressive procedures) and anxiety.
According to James (2013), in the USA, up to 440 000 patients die per year due to human error of medical specialists. One direct solution to this is to resort to a second specialist to review all the decisions made. But the associated cost is too high, that's why the CAD systems can be an alternative solution, since they can directly help the medical specialist when making diagnostics at a decreased cost. In the area of breast imaging, Moura and Guevara López (2013) and Zhang et al. (2011) have shown that CAD systems can help decreasing the number of false negatives while not decreasing the number of true positives.

## 1.2 Objectives

The purpose of our work is to improve the classification of tumors shown in mammograms, based on the results reported by Augusto (2014) with Multiple Kernel Learning (MKL) and the Breast

Cancer Digital Repository (BCDR) data. We observed that MKL as applied by Augusto (2014) did not take into account that the class distribution can be imbalanced, therefore, we found an opportunity for improvements. Our methodology, besides using a modified form of MKL, adjusts the learning using weights associated with the classes. In this work we also present and study all the steps necessary to theoretically detect regions of interest (ROI), extract features and classify ROI automatically.

## 1.3 Similar Works

Over the years, there have been several works in the study of mammograms. The analysis of a mammogram takes several phases. The first thing a radiologist does when analysing an image is to ROI. Various computer-aided systems have been built to automatically detect ROI in images, works done by Rahmati and Ayatollahi (2009), Melouah and Merouani (2008) and Oliver et al. (2008) achieved up to 82% correct classification of regions. After detecting the ROI, the next step is to automatically extract features from it, this was done by Kitanovski et al. (2011). Finally, algorithms are applied to those features in order to classify the region as malignant or benign, this final step is the focus of this work. Arguably, Support Vector Machines are the most common method used to classify mammograms. Mammogram classification works like El-Naqa et al. (2002) and El-Naqa et al. (2004) show good results in the ROC and precision-recall curves. A study by Wei et al. (2005) shows that kernel methods achieve better results than others models, also Wei et al. (2009) achieved 82% accuracy on the classification of micro-calcifications. Several works tried to apply MKL to this domain, but did not present their methodology in detail, and report accuracies above 98% like in Zare et al. (2014) and Yang et al. (2013). Some other works, also using MKL, present more realistic results (Ma et al. (2015)), reporting AUC of 85%, although not providing much detail on the methodology. Some works like Liu et al. (2012) focused only in classifying if a region is a mass or simply normal tissue, using SVM with Radial Basis Function (RBF) kernel for classification obtaining 86.6% accuracy value. Finally the work of Augusto (2014), on which our work is based, does classification of all the types ROI using MKL, and is able to achieve a ROC curve AUC value of 87%.

## 1.4 Scientific Challenges

To achieve the results shown in Chapter 4 we had to solve several problems related with the specific type of data we wanted to use and implementation of the model itself.
One of the most important challenges we face is the class distribution. Fortunately, in this domain, the number of malignant cases is much smaller than the number of benign cases, but this imbalance causes difficulties to automatic classifiers, mainly because almost every model will give more weight to the more prevalent class. A simple example: if our dataset has one hundred benign objects and just 5 malignant objects, during the training most of the classifiers will have to ignore some of these objects to get better results, but ignoring 10 benign objects has

a different impact than ignoring 4 malignant. To address this problem we followed a strategy that allows the use of weights for each class.
In mammograms, expert radiologists usually perform manual annotation of findings that they detect in the images. These annotation are described in Section 2.1.2. We use these annotations as features. These annotations represent just a small part of all the features used, and we call them clinical information. The other big group of features in our datasets is extracted from the mammogram image. These features are described in Section 2.3 and most of them can not be directly related with the Breast imaging-reporting and data system (BI-RADS) medical system. We need datasets with already extracted features from the mammogram image that also contain clinical data, this type of data is not very common due to its specificity. Because of that we had to extract the features from the Digital Database for Screening Mammography (DDSM) that contains mammograms and also the clinical data annotated by a specialist.
Since our data is unbalanced, an MKL with weights is necessary. The MKL uses internally the same solver as an SVM, and so it is easy to just change the original solver for one with weights like we did.
In order to perform a fair comparison between the datasets, the set of features of all datasets should coincide. To achieve that, we implemented a feature extractor for DDSM.
For the creation of the final model several parameters from the MKL must be tuned while achieving an algorithm that should not allow overfitting.

## 1.5 Contributions

We can summarize our contributions below:

- Review of the state-of-the-art on breast image classification using SVM-based methods.
- Implementation of a methodology based on Cost-Sensitive SVM and Non-Linear MKL.
- Application of our methodology to breast image classification.
- Improvement of results over other methods based on SVM and MKL.

## 1.6 Structure

This work is divided in five main chapters. In Chapter 2 we suggest and explain all the technologies, areas and terminologies used in our work, it is written with the intuit of teaching the reader, and allow the recreation of the work done. During the explanations we cite all the papers that we used for reaching our final results. In Chapter 3 we show all the works that recently have been published in the area of data mining applied to mammogram images, we describe and divide in groups so that the reader can easily find the papers that represent the state of the art in a specific topic. In Chapter 4 all the methods, algorithms and materials are

described. It is also here that we show and discuss results. Finally in Chapter 6 we express our conclusions by looking at the results and comparing them with other work, we also describe what we expect to be the future after this work, and what should be achieved with it.

# Chapter 2

# Fundamental Concepts

In this chapter we explain all the technologies and techniques that were necessary to achieve this work. It is divided by degree of the mathematical and algorithm complexity. Starts by the discussion of health-related topics and finishes with the explanation of the SVM and MKL algorithm. We also explain many terms used in this paper that may not be understood by readers outside the area of machine learning.

## 2.1 Breast Cancer

### 2.1.1 Diagnosis

The most common way to diagnose breast cancer is by using mammograms like the one in the Figure 2.1, they are the cheapest least invasive technique and easy to obtain. The problem is that the process of detecting masses and classifying them is done by a health professional, and the process of analysing these images can be very tiring and difficult. Sometimes two different professionals take on the same mammogram with the purpose of reducing error, but even so it's very difficult to obtain completely certain diagnosis. Because of that and for the safety of the patient over-diagnosis and over-treatment is performed, usually leading patients to more invasive interventions like biopsies or excision.

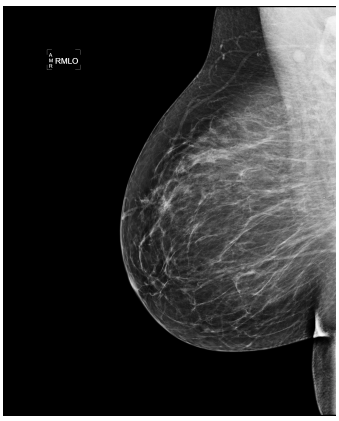



Figure 2.1: Generic Mammogram.

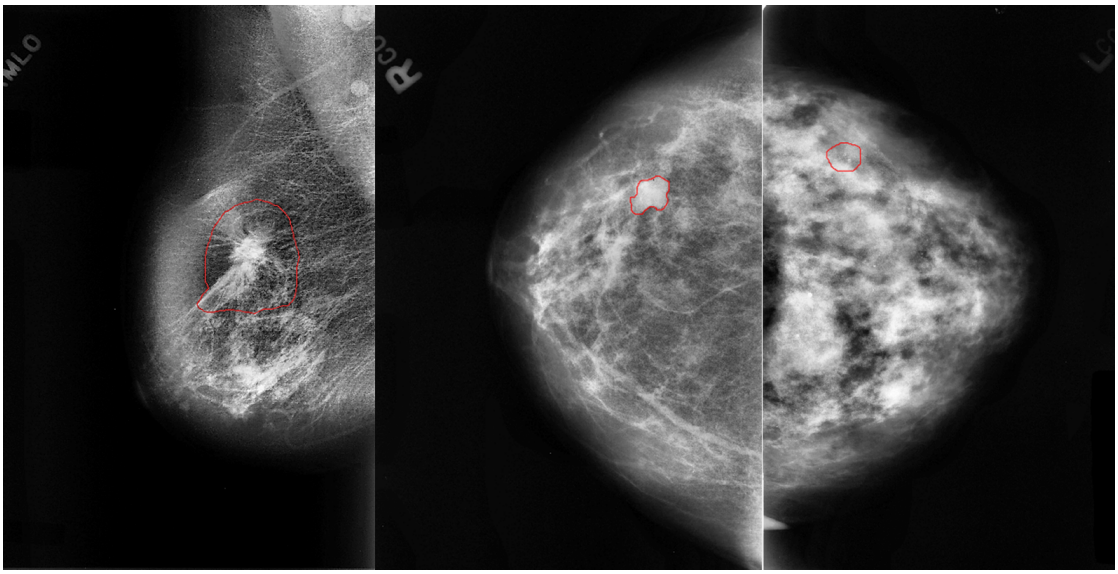


Figure 2.2: Three mammograms containing an Architectural Distortion, Mass and Calcifications.

### 2.1.2 BI-RADS

BI-RADS (D'Orsi (2013)) is the standard for Breast Image Reports. It is used not only for Mammograms but also for Ultrasounds, Magnetic Resonances and used to describe category of the tumour. BI-RADS attributes are mainly all types of findings that can appear in a mammogram, we followed the original reference card for BI-RADS that can be found in the American College of Radiology (ACR), and made an interpretation of each type of findings and features according to Kopans (2007). Examples of mammograms showing distinct findings circled in red, are shown in Figure 2.2

· **Masses** can be detected by external touch and are made of breast tissue or cysts due to the accumulation of fluids. A mass is described by its shape, margin, density and size.

· **Calcifications** are accumulations of calcium within the breast, usually they are product of lesions, inflammations and age. They are divided in two distinct groups, the typically benign and the ones of suspicious morphology that may be malignant. They can be found in clusters or alone, because of that the type of distribution must be described.

· **Architectural Distortion** is a region that shows abnormal arrangement of breast tissue, often a radial or perhaps a somewhat random pattern, but without any associated mass or calcification acting as apparent cause of this distortion.

· **Skin lesions** are not important for breast cancer detection but can be seen in the mammogram. They must be noted to avoid wrong conclusions, like misclassify them as masses or calcifications.

· **Breast Asymmetries** happen when an area of breast tissue in one breast side is different

from the same area in the other breast. This helps in the process of detecting other abnormalities and malignancy.

· **Intra mammary lymph nodes** are by themselves not malignant but they are the primary sites of metastases[1], because of that they have clinical significance and should be noted.

· **Solitary dilated duct** are very important for malignancy detection because even though that when alone they are benign, it is known that when in combination with other symptoms like masses or distortions, they can be the deciding factor to confirm the existence of malignancy.

· **Associated Features** are more general and usually less relevant clinical features, like skin retraction, nipple retraction or skin thickening.

## 2.2 Image Processing

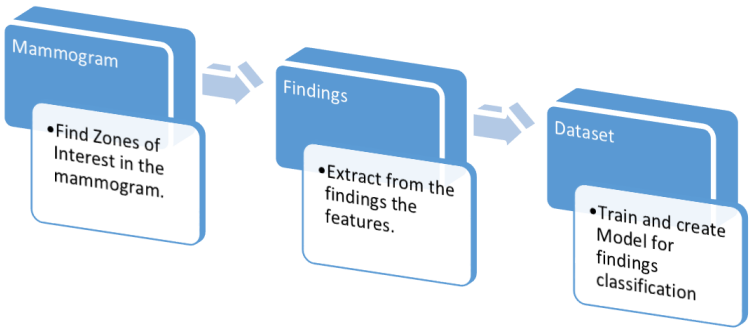



Figure 2.3: Image Analysis Flow

Image processing usually follows three main phases: segmentation, feature extraction and, possibly, classification. The flow of image analysis is shown in Figure 2.3. The first task is the detection of the ROI in the mammogram, we can also call them findings or anomalies, this process is called Segmentation and is explained in Section 2.2.1. Features are extracted from these findings, the type of features and how they are extracted is explained in Section 2.3. All the features must be ranked by level of interest to the model, a feature is interesting when it increases the performance of the model. The ranking of the features is done by a feature selector that is described in Section 2.3.2. Finally by using the clinical and extracted features the object that describes the finding should be classified as malignant or benign, this is done by the MKL model that is described in Section 2.6.1.4. Each of these tasks have their own problems and difficulties associated, and our work will focus in the feature extraction, feature selection and object classification tasks.

[1]Metastases is the name given when malignant cells are released from the tumor and spread to the rest of the body.

### 2.2.1 Image Segmentation

Image segmentation is the process of partitioning an image (in our work, a mammogram) into multiple segments. This allows the detection of ROI. In our work, this process was done manually by specialists, but it can be done automatically, and usually this process is divided into three distinct phases.

- **Image Treatment** According to Khan et al. (2015) before doing any type of partitioning the image should be reduced to only the breast. To achieve that the image must go through some transformations. The black or white borders should be removed. This can be done by looking at the intensity of the pixels and remove all those that are near 0 or 1. A mammogram of the left and right breast is usually mirrored. Because of that we must orient the breasts all to the same side, this is necessary because we need all the images with the breast on the same side to make the implementation and process of segmentation simpler. The removal of artifacts can be easily done by finding the breast boundary with the Otsu threshold method described by Wenqin (1993), and setting everything outside to black. Then an average filter or other can be applied to the image but that is dependent of the type of image we are working with. (A thresholding method replaces each pixel in an image with a black pixel if the intensity is less than some value F, or white pixel if the intensity is greater than F. The challenge in this process is to find the value of F.)
- **Removal of Chest Muscle** To obtain just the breast region the chest muscle must be removed because this muscle tissue looks similar to the one of malignant ROI. There are some different techniques that can be applied to do that. Khan et al. (2015) uses a threshold to find the border between breast and muscle. Qayyum and Basit (2016) uses a canny edge detection to create a line between breast and muscle. Kowsalya and Priyaa (2016a) remove the borders of the image removing this way most of the muscle area.
- **Detecting regions of interest** The automatic detection of ROI can be done by applying thresholds that allow the finding of the boundaries of each ROI, also it can be done by using the snakes algorithm like the one described by Yuen et al. (1996) that adjusts a line or circle to the nearest ROI, this is done by Chakraborty et al. (2016).

### 2.2.2 Extracted Features

There are at least 3 big groups of features that can be extracted from a black and white image, each one is able to obtain different knowledge about the image. According to Moura and Guevara López (2013) these features are able to describe the light or intensity, texture, Shape and Location.

**Intensity descriptors** are calculated using the gray levels of the pixels. This group of features describes the level of luminosity of a region of interest. It uses the Standard deviation, Minimum, Average, Median and Mode of the intensity values.

**Texture descriptors** are extracted using a gray-level co-occurrences matrix. In Figure 2.4, left matrix, we have the representation of gray levels of an image with 4x5 pixels, whose minimum gray level is 1 and maximum is 8. This will produce a co-occurence matrix of size 8x8. Starting from left-to-right, row-wise in the matrix on the left, we start filling up the values of the co-ocurrence matrix, in the right. For example, the first pair of gray levels 1,1 appears only once in the left matrix, therefore, entry 1,1 of the co-occurrence matrix will be filled up with 1. The pair 1,5 also appears only once, which fills up entry 1,5 with the value 1. Pair 1,2 appears twice, so entry 1,2 will be filled up with the value 2, and so on, and so forth. The gray-level co-occurrence matrix (GLCM) is used then to find adjacent areas of the picture that are more relevant or less relevant regarding the gray levels. We will use $P(x, y)$ as the probability of pixels with gray-level x occurring together to pixels with gray-level y. $P_+(i)$ or $P_-(i)$ are the sum or subtracted probability of two co-occurrence matrix coordinates $i = x + y$. L=8 will be maximum of the gray levels, M is the mean of all the P(x,y), $i_-$=|x-y| and $i_+$=x+y.

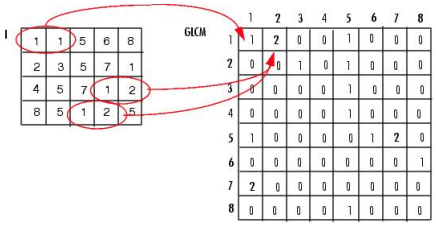



Figure 2.4: **The matrix on the left represents the image with the gray levels of the pixels, on the right is the gray-level co-occurrence matrix.**

- **Energy** $\sum_{x=1}^{L} \sum_{y=1}^{L} P(x, y)^2$
- **Homogeneity** $\sum_{x=1}^{L} \sum_{y=1}^{L} \frac{P(x, y)}{1 + (x - y)^2}$
- **Contrast** $\sum_{x=1}^{L} \sum_{y=1}^{L} (x - y)^2 P(x, y)$
- **Variance** $\sum_{x=1}^{L} \sum_{y=1}^{L} (x - M)^2 P(x, y)$
- **Entropy** $-1 * \sum_{x=1}^{L} \sum_{y=1}^{L} P(x, y) * (log(P(x, y)))$
- **Sum Average** $\sum_{i=2}^{2*L} i * P_+(i)$
- **Sum Entropy** $-1 * \sum_{i=2}^{2*L} P_+(i) * (log(P_+(i)))$
- **Sum Variance** $\sum_{i=2}^{2*L} (i - SumEntropy)^2 p_+(i)$
- **Difference Entropy** $-1 * \sum_{i=2}^{2*L} P_-(i) * (log(P_-(i)))$

**Shape and Location descriptors** is the last group and the one that provides information about the geometrical characteristics of the region of interest. It contains the following features:

- **Perimeter** Number of pixels in the edge of the segment.
- **Area** Number of pixels inside the segment.
- **Circularity** $4\pi * \frac{area}{perimeter^2}$

- **Elongation** $\frac{X}{Y}$ Where X is the minor axis and Y the major axis of the ellipse that encloses our region of interest.
- **Y centroid** $\frac{min(Yaxis) + max(Yaxis)}{2}$.
- **X centroid** $\frac{min(Xaxis) + max(Xaxis)}{2}$.
- **X & Y Centroid** Normalized coordinates of the centre of the set of pixels that belong to the segmented lesion.
- **Solidity** $\frac{area}{|H|}$ Where |H| is the total pixels of the convex hull of the segmented region.

## 2.3 Features

### 2.3.1 Feature Types

According to Berthold et al. (2010) features can be categorized in 4 different types.

- **Continuous Numeric Features** are numbers and have a value. If we define a range between two possible values of feature, there is an infinite number of different values that belong to that interval, and since there is no interval between any two values we call it continuous. This is more common in features that are results from formulas since there is no theoretical limit to how much precise a number can be.
- **Discrete Numeric Features** are numbers and have a value, If we define a range between two possible values of feature, there are zero or a finite number of possible different values in it. Examples of Discrete features can be the number of days in a year or the time using only hours, minutes and seconds.
- **Ordinal Features** are features that do not have a value but have an order and usually there is a predefined number of different words that belong to the feature. An example of this can be a feature that uses Very low, Low, Normal, High and Very high as values.
- **Categorical Features** do not have order and do not represent any numerical value, they are simply lists of objects. Examples of these can be the name of different fruits in a basket or simply a True False (Binary) feature. In our context it could be the presence or absence of a mass or calcification.

### 2.3.2 Feature selection and UFilter

Feature selection is the process of selecting or ranking features by level of interest to the model, it allows the filtering of features that may decrease the performance or are irrelevant. One of the

main feature selectors we use is the Ufilter created by Pérez et al. (2015). It is a feature selector oriented for *breast cancer diagnosis on mammography*, inspired by the U test of Mann-Whitney (described in Nachar et al. (2008)) that calculates the difference between data samples of a feature. For this problem, sample is the group of objects that belong to a class. In a binary class problem each feature will have two samples one for the negative and one for the positive class. For better understanding we will be explaining the algorithm just for a binary class.
Let $F = f_1, f_2, ..., f_t$ be a set of features where $t$ is the total number of features, and let $f_i = v_1, v_2 ..., v_N$ be a set of values for feature $f_i$, where $N$ is the total number of values of feature $i$. Ufilter orders the values from a feature and solves ties by averaging the positions of the tied objects when other features are ordered. Then the sum of the positions of the objects of each class are stored in $S_M$ and in $S_B$ ($B$ is benign and $M$ is malignant):

$$\begin{aligned} S_B &= \sum_{j \in B} Pos_{f_j} \\ S_M &= \sum_{j \in M} Pos_{f_j} \end{aligned} \tag{2.1}$$

In the equations of 2.1 $B$ and $M$ are the groups of features which class is benign or malignant, the variable $Pos_{f_j}$ is the position value of the feature $j$. Using the number of malignant and benign ($N_M$ and $N_B$) and the sum of the respective positions ($S_M$ and $S_B$) we can obtain the $\upsilon$-values (2.2) and the Z-values (2.3).

$$\begin{aligned} \upsilon_B &= N_B * N_M + \frac{N_B(N_B+1)}{2} - S_B \\ \upsilon_M &= N_B * N_M + \frac{N_M(N_M+1)}{2} - S_M \end{aligned} \tag{2.2}$$

These equations are the same as the ones from Nachar et al. (2008). The following equation 2.3 is where both models differ. In the original we would select the minimum from both of $\upsilon$-values to calculate equation 2.3 and accept or reject the null hypothesis at a given level of significance $\alpha = 0.05$. In the uFilter method it is computed both of the Z-indicators (one for each class).

$$\begin{aligned} Z_B &= \frac{\upsilon_B - \nu}{\sigma_\nu} \\ Z_M &= \frac{\upsilon_M - \nu}{\sigma_\nu} \end{aligned} \tag{2.3}$$

Where $\nu$ is the mean and $\sigma_\nu$ is the standard deviation. The score of each feature is given by 2.4. This score is the total difference between the results obtained by the samples made of each class.

$$W_i = |Z_B - Z_M| \tag{2.4}$$

## 2.4 Performance Metrics

### 2.4.1 Confusion Matrix

Table 2.1: Confusion Matrix

| | Predicted Positive | Predicted Negative | Totals | |
|---|---|---|---|---|
| Positive Condition | True Positive | False Negative | $Condition_P$ | Rec all $\frac{TP}{Condition_P}$ |
| Negative Condition | False Positive | True Negative | $Condition_N$ | Specificity $\frac{TN}{Condition_N}$ |
| Totals | $Predicted_P$ | $Predicted_N$ | | |
| Accuracy $\frac{TP+TN}{TP+FN+FP+TN}$ | Precision $\frac{TP}{Predicted_P}$ | False Omission Rate $\frac{FN}{Predicted_N}$ | F1Score $\frac{}{(1+\beta^2)*\frac{PR*REC}{(\beta^2*PR)+REC}}$ | |

To describe results and compare models several metrics are used, all of them depend on the data that is stored in a confusion matrix like the one in Table 2.1. This is used for binary classification problems and the matrix is built by counting the number of times the model correctly or not classifies the objects. Predicted positive or negative refers to the prediction of the algorithm, and Condition positive or negative is the actual value of the class. True positive or negative is one object that has been correctly classified, False positive or negative is one object that has been wrongly classified. We use almost all of the metrics in Table 2.1. Accuracy is the easiest to understand because it simply shows the percentage of objects that are correctly predicted. Recall or Specificity use only objects from one class and shows the percentage of correctly classified objects for that class, precision or FOR is similar but uses only objects that have been classified with the same class and for those, show the percentage that have been correctly classified. All these different metrics should be used in combination to better explain the results obtained by a model.

### 2.4.2 Curves

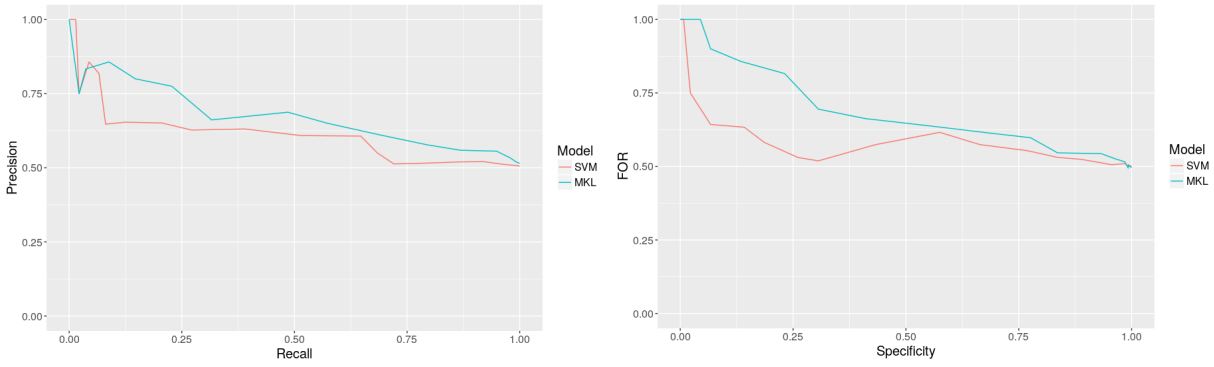



Figure 2.5: Example of Two Different Curves

There are several types of curves that can be built using the metrics from the confusion matrix.

In our work we use three. The receiver operating characteristic ROC curve, the precision-recall and the FOR-specificity curve.
Imagine a binary classification problem, whose model can output a probability for the binary class, for example the object $x_i$ got 0.6. There is a 60% chance that the object $x_i$ is positive. When we have this kind of results it is possible to create thresholds that make all objects above 50% positive and the ones below negative, the same can be done for 40%, 30% and so on. We create a confusion matrix for each threshold and then use them to create curves that describe the performance of the model. To create an ROC curve we map all the $TruePositive$ in $y$ axis and $FalsePositives$ in the $x$ axis, the precision-recall curve plots *precision* in $y$ axis and *recall* in the $x$ axis. From these curves we can calculate the Area Under The Curve. AUC is one of the most common used metrics in data mining because it ranges between one and zero, and combines two different metrics that compose the curve.

### 2.4.3 Unbalanced Data Metrics

Table 2.2: Sample Confusion Matrix

| Predited / Condition | Positive | Negative | |
|---|---|---|---|
| Positive | 1 | 3 | Recall 25% |
| Negative | 4 | 16 | Specificity 80% |
| Accuracy 71% | Precision 20% | FOR 84% | F1Score 22% |

When dealing with unbalanced datasets the use of some metrics can result in very wrong conclusions this is a known problem in literature, Provost et al. (1997) and Gu et al. (2009) did a study on this, and concluded that accuracy and other metrics may not provide accurate measures of the classification performance of imbalanced data sets.
We will be using an example assuming that we have a dataset of 20 Negative and 4 Positive objects, and that our confusion matrix is Table 2.2. The best way to explain why some metrics will not work is by looking at the recall and specificity. We can compare how well each of the classes is being classified, and in the example, that the difference in results for both classes is 65%. Precision and FOR shows the difference in the weight of a misclassification for both classes, the difference between both class results is 64%. By looking at these four metrics it is possible to conclude that this model is doing a bad job in the classification of the Positive class. The problem arises when we know that our model only classifies correctly one quarter of the positive class (recall), but we present only the accuracy of the model. This way it's is possible to hide from the reader how bad it is, since 71% can be an acceptable value in many types of problems. The same happens with any of the metrics that only represent the negative class or do an average of both classes.
The ROC AUC curve can also suffer from this problem and because of that not only we also show the ROC AUC curve for the negative class but we also use precision-recall and FOR-specificity curves.

### 2.4.4 P-Value

According to Westfall and Young (1993) the P-Value allows us to compare the results from two different models and understand if these are statistically different, we can assume that the results are different when P-Value is inferior to 0.05. Let us assume we have experimented two classification models in a dataset of 200 objects of ROI, the Model A will be considered to contain the Observed values and Model B the expected Values. First step for the calculation of the P-Value is the definition of the degrees of freedom. The number of degrees of freedom is equivalent to the number of categories minus one, we have malignant and benign classes so our Degree of freedom is 1. We have to calculate the Chi-Square $\chi^2 = \sum_{i=1}^{N}((o_i - e_i)^2/e_i)$ where N is the number of categories, $o$ and $e$ are the observed and expected values for category $i$. Finally just to use a $\chi^2$ table and find in the row corresponding to our degree of freedom the value that is nearest to ours and the corresponding P-Value.

## 2.5 Model Validation

To evaluate the performance of a model we can simply cut 20% of the dataset as a test set and leave the remaining 80% as a training set. This strategy can lead to overfitting of the model. Overfitting occurs when the model follows the training and test datasets very rigorously, which leads to a high loss of performance on more generalized data. This can be avoided by applying Bootstrap or Cross-Validation methods. According to Berthold et al. (2010) bootstrap consists in sampling $n$ objects from the dataset with reposition several times, creating this way $x$ datasets that can be used for experiments. Other algorithm is cross validation, it divides the dataset in $n$ folds that can be used for experiments. If the dataset is very small Leave one Out method can be used, it consists in doing the training several times while leaving one object out. Bellow is the pseudo code for each of these methods.

```
1 begin
2     train = sample(dataset);
3     test = dataset - train ;
4     Results = Experiment(train,test);
5 end
```

**Algorithm 1:** Simple Train Test Method

```
1 begin
2     for 1 to Number_of_Samples do
3         bootdata = bootsample(dataset);
4         train = sample(data);
5         test = dataset - train ;
6         Results.add(Experiment(train,test));
7     end
8     Final_Results= Average(Results);
9 end
```

**Algorithm 2:** Bootstrap

```
1  begin
2      F_Array=create_K_folds(Data,K);
3      forall fold of F_Array do
4          train=F_Array - fold;
5          test=fold;
6          CMatrix = Experiment(train,test);
7          Results.add(CMatrix);
8      end
9      Final_Results= Measures(Results);
10 end
```

**Algorithm 3:** K-Fold Cross Validation

```
1 begin
2     forall object of Dataset do
3         train=F_Array - object;
4         test=object;
5         CMatrix = Experiment(train,test);
6         Results.add(CMatrix);
7     end
8     Final_Results= Measures(Results);
9 end
```

**Algorithm 4:** Leave One Out

## 2.6 Machine Learning

In this Section we will be describing the state of the art algorithms. Because our work is based in the use of MKL and SVM, those are the two algorithms we will be giving more focus.

### 2.6.1 Methods for Classification

The following Sections should give some insight on what and how data mining algorithms can be used for classification problems. SVM and MKL will be described with more detail in another Section.

#### 2.6.1.1 Decision Trees

Decision trees are the simplest method that can be used for classification problems, like the name suggests the algorithm creates a tree similar to the Figure 2.6. The internal nodes represent the conditions for splitting the data and the leaves the final classification. The classification of one object is done by following the corresponding path from the root to the leaves. The construction of the tree can be done by Algorithm 5 suggested by Breiman et al. (1984), this algorithm is able to create a big tree that is not good, because by having too many nodes we end up with a complex and difficult to understand tree, also this can lead to overfitting of the model. The process of reducing the size of the tree is called pruning, and one way to do this is by iterating through the nodes (bottom up) starting at the leaves and measure if a parent of a node contains the same information as the son. When that happens we remove the sons and leave the parent node with the class that was more dominant on the sons. One of the biggest benefits of using decision trees is that results are easy to explain by looking at the structure of the tree, because each of the splits can be taken as a logic and explainable reason for the final classification.

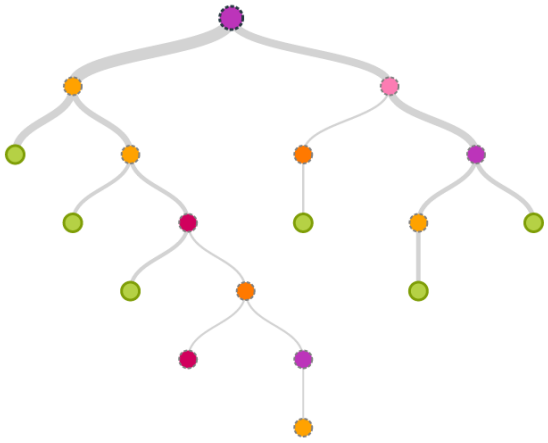


Figure 2.6: Decision Tree Structure

```
1  Algorithm(BuildDecisionTree(𝒟, 𝒜));
   input : Training Data 𝒟 with 𝒜 Attributes
2  begin
3      if all elements in 𝒟 belong to one class then
4          return node with class label;
5      end
6      else if 𝒜 = ∅ then
7          return node with majority class label in 𝒟;
8      end
9      else
10         select attribute A ∈ 𝒜 which best classifies 𝒟;
11         create new node holding decision attribute A;
12         for each split of A do
13             add new branch create a new 𝒟_{ν_A} ∈ 𝒟 for which split condition holds;
14             if 𝒟_{ν_A} = ∅ then
15                 return node with majority class label in 𝒟;
16                 else
17                     add subtree returned by calling BuildDecisionTree(𝒟_{ν_A}, (𝒜\A));
18                 end
19             end
20         end
21         return node;
22     end
23 end
```

**Algorithm 5:** BuildDecisionTree($\mathcal{D}, \mathcal{A}$)

#### 2.6.1.2 Artificial Neural Networks

Artificial Neural Networks (ANN) try to recreate the way that the neural network of brain works, the idea is to make information travel a network like the one in Figure 2.7. The inside of this structure is built of functions that create a flow of data, these functions are connected by weights that change each time the ANN is used, this is one of the best characteristics of ANN. ANN and SVM unlike the other algorithms described in this work, produce results that cannot be explained using features values and or conditions. For more information on the algorithm we propose the reading of the Section about ANN of Berthold et al. (2010).

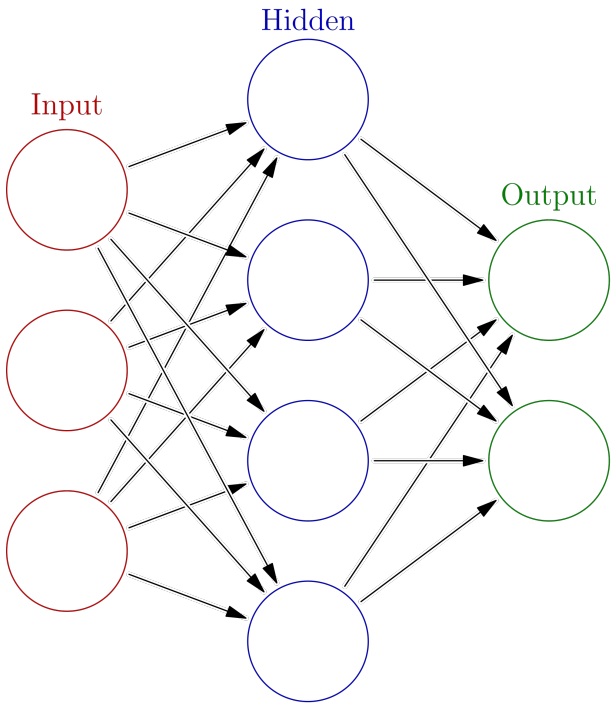



Figure 2.7: Artificial Neural Networks Structure

#### 2.6.1.3 Bayes Classifiers

Bayesian-Network (BN) present a network of connected nodes that can be seen in Figure 2.8 where each node contains the probability of attribute X. According to Berthold et al. (2010), BN are built around the probabilities of several attributes. In BN we predict the probability of an object $\mathbf{x}$ belong to a class y. In Equation 2.5 and the following we use $P$ as the probability, $pred$ as the prediction.

$$pred(\mathbf{x}) = arg_{y \in dom(Y)} \ max \ P(y|\mathbf{x}) \tag{2.5}$$

This leads us to the Bayes Theorem (eq.2.6):

$$pred(\mathbf{x}) = arg_{y \in dom(Y)} \ max \ \frac{P(\mathbf{x}|y)P(y)}{P(\mathbf{x})} \tag{2.6}$$

This type of classifier, like the decision tree, achieves results that are easy to explain, because we can use the probabilities of the network and make conclusions of what can be the most important factors in the classification and create an explainable logic of factors. There are more than one type of BN and there are several ways to build BN, for more details on this we recommend reading Cowell et al. (2006) Sections 2 to 4 for a more introductory approach.

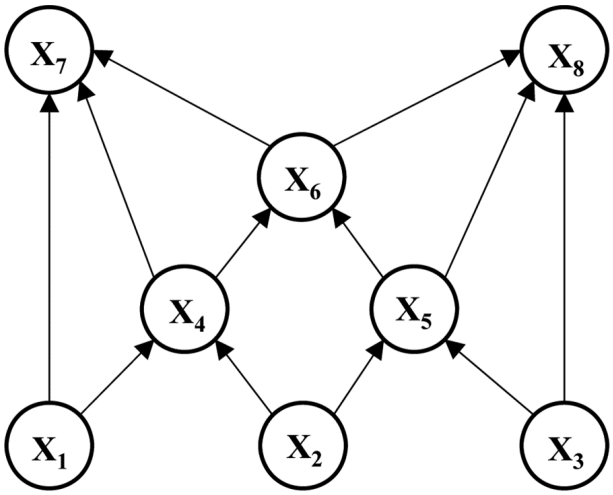



Figure 2.8: Bayesian-Network Structure

#### 2.6.1.4 Support Vector Machines

An SVM takes a sample of N objects $\{x_i, y_i\}_{i=1}^{N}$, where $x_i$ is a feature vector that describes the class $y_i \in \{+1, -1\}$ of object $i$, and finds a way to separate the objects in two classes. This is done by using what we call a hyperplane (equation 2.7).

$$f(x_i) = w * x_i + b \tag{2.7}$$

In the hyperplane 2.7, $w$ and $x$ are vectors. Vector $w$ is orthogonal to the hyperplane and $x_i$ is the vector corresponding to the object $i$ we want to classify. Applying $f$ to $x_i$ results on the classification of $x_i$, which, most of the time, should agree with $y_i$, in order to reduce the classification error. According to Bishop (2006); Chang and Lin (2011); Gonen and Alpayd (2011), the SVM training can be reduced to the optimization problem (Equation 2.8).

$$\begin{aligned} \underset{w,\zeta,b}{\text{minimize}} \quad & \frac{1}{2} w^\top w + C \sum_{i=1}^{N} \zeta_i \\ \text{subject to} \quad & y_i * (\langle w, \phi(x_i) \rangle + b) \geq 1 - \zeta_i \qquad 1 \leq i \leq N \end{aligned} \tag{2.8}$$

Where $w$ is the vector we want to find, $\phi(x_i)$ is the vector $x_i$ projected to another dimension (this is done by a so-called Kernel that we will be describing later), $\zeta$ is a set of slack values used to find an optimal hyperplane (when a slack value is big enough, the SVM can leave a vector on the wrong side of the hyperplane misclassification is allowed in order not to overfit the model), $C$ is according to Ben-Hur and Weston (2010), the smoothing factor, a value that maintains the balance between the minimization of $w$ and $\zeta$, this means that the bigger the $C$ the smaller the margins will be because it increases the number of support vectors, and finally $N$ is the number of support vectors (this will be discuss later). In the domain of breast cancer, where the learning task is to discriminate between malignant and benign cases, fortunately, the number of malignant cases is much smaller than the number of benign, but this leads to a well-known classification

problem: learning from unbalanced number of elements per class. Masnadi-Shirazi et al. (2012) studied the problem and proposed a Cost-Sensitive Support Vector Machine (CSSVM) to handle imbalanced classes, the cost-sensitive optimization problem is the following:

$$\begin{aligned} \underset{w,\zeta,b}{\text{minimize}} \quad & \frac{1}{2}w^\top w + C[C_1 \sum_{i \in y_i=+1}^{n} \zeta_i + \frac{1}{k} \sum_{i \in y_i=-1}^{n} \zeta_i] \\ \text{subject to} \quad & \langle w, \phi(x_i)\rangle + b \geq 1 - \zeta_i \qquad y_i = +1 \\ & \langle w, \phi(x_i)\rangle + b \leq -k + \zeta_i \qquad y_i = -1 \\ \text{with} \quad & k = \frac{1}{2C_{-1} - 1} \qquad 0 \leq k \leq 1 \leq \frac{1}{k} \leq C_1 \end{aligned} \tag{2.9}$$

In this equation $y_i$ is removed from $y_i * (\langle w, \phi(x_i)\rangle + b)$ and because of that we have to split the equation, and create one for each class margin. This allows the use of 3 new variables. $k$ imposes a smaller margin on negative objects when the data is separable, $C_1$ and $C_{-1}$ are weights for each class because they directly increase or decrease the slack values $\zeta$ of each class. $C_{-1}$ controls the difference in the size of the margins, which means that the bigger the $C_{-1}$ the smaller the margin to the negative class, this happens because $k$ decreases and the margin of the negative class is dependent of the size of k $\langle w, \phi(x_i)\rangle + b \leq -k$. If we fix this value and increase $C_1$ knowing that $C_1 > 2 * C_{-1} - 1$ we increase cost on the error of the positive class by increasing the cost of the slack values. This allows us to create a model where we can guarantee that the error rate of both classes can be minimized without big losses of accuracy. To solve any of the inequations in 2.9, we first need to apply the Lagrangian dual function to obtain the dual problem for the function 2.8 (cf. 2.10).

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{N} a_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j y_i y_j K(x_i, x_j) \\ \text{subject to} \quad & \sum_{i=1}^{N} a_i y_i = 0 \qquad 0 \leq a_i \leq C \qquad 1 \leq i \leq N \end{aligned} \tag{2.10}$$

And for the function 2.9 (cf. 2.11).

$$\begin{aligned} \text{maximize} & \sum_{i=1}^{N} a_i (\frac{y_i + 1}{2} - \frac{k(y_i - 1)}{2}) \quad -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j y_i y_j K(x_i, x_j) \\ \text{subject to} & \sum_{i=1}^{N} a_i y_i = 0 \qquad 1 \leq i \leq N \\ & 0 \leq a_i \leq C C_1 \qquad y_i = 1 \\ & 0 \leq a_i \leq \frac{C}{k} \qquad y_i = -1 \end{aligned} \tag{2.11}$$

In both equations, 2.10 and 2.11, the kernel formula $K(x_i, x_j)$, is equivalent to $\phi(x_i)\phi(x)$. By solving these equations we can obtain 2.12.

$$w = \sum_{i=1}^{N} y_i a_i \phi(x_i) \tag{2.12}$$

If we look again at the initial function of the hyperplane 2.7, it is easy to see that 2.13 is true.

$$w * x + b = \sum_{i=1}^{N} y_i a_i \phi(x_i)\phi(x) + b \tag{2.13}$$

In 2.13, $x$ is the new object we want to classify and $x_i$ our support vectors. Support vectors are the objects that make the margins of our hyperplane, and correspond to data points that are on the edge of each class. When training an SVM, we only need the support vectors, which reduces the complexity of the SVM training by confining the search region to a small area. One problem that an SVM does not solve is when data is not linearly separable. If data is not linearly separable, it may be necessary to project the data to another dimension, and this is done using Kernels. A kernel is simply an equation that multiplies 2 vectors, but instead of doing a normal multiplication it does it while increments the dimensions of the vectors, without increasing the complexity of the problem. For example, if we apply the polynomial kernel $\phi(x)\phi(y) = (x \top y + c)^d$, this will increase the distance between the objects and allow the SVM to better fit the hyperplane. Several kernels can be used by an SVM. We selected the Heavy tailed radial basis function kernel (HTRBF) that appears originally in Hou et al. (2011); Chapelle et al. (1999) and is shown in 2.14. This kernel has better performance than more commonly used kernels, and it allows the creation of different kernel versions by changing the parameters $a$, $b$ and $\gamma$. We also applied the Gaussian (2.15) and Anova (2.16) Kernels due to their common use in SVMs.

Heavy tailed radial basis function

$$\phi(x)\phi(y) = e^{-\gamma||x_i^a - y_i^a||^b} \tag{2.14}$$

Gaussian Kernel

$$\phi(x)\phi(y) = e^{-\frac{||x_i - y_i||^2}{2\alpha^2}} \tag{2.15}$$

Anova Kernel

$$\phi(x)\phi(y) = \sum_{k=1}^{L} e^{-\alpha(x_i^k - y_i^k)^2)^d} \tag{2.16}$$

We handle all features equally, applying the same method and, if needed, the same kernel to all of them. However not all features are of the same type. They can be continuous, discrete, nominal or binary, and even when they fall under the same category, the values distribution can be different. Therefore, each one would need a specific kernel to help maximizing the separation of the classes.

#### 2.6.1.5 Multiple Kernel Learning

MKL allows the use of multiple kernels, several applied to each variable. It then creates a Linear or Non-Linear combination of kernels and tries to find the weights $\eta$ for this equation (2.17). We chose to improve this particular strategy because of the previous work done by Augusto (2014)

that shows good results by using this kernel combination technique.

$$K(x,y) = \sum_{i=1}^{P} \eta_i k_i(x,y) \tag{2.17}$$

In Equation (2.17), $P$ is the number of different kernels. We chose Non-Linear MKL (NLMKL) to find the weights of $\eta$ and according to Gonen and Alpayd (2011), this implementation has shown the best results when compared with other MKL implementations and with SVM with simple kernels. This implementation also allowed us to use a non linear kernel combination:

$$K(x,y) = \sum_{i,j=1}^{P} \eta_j k_j(x,y) \eta_i k_i(x,y) \tag{2.18}$$

The NLMKL is inspired by the standard kernel ridge regression (KRR) dual optimization algorithm for a fixed kernel matrix, done by Saunders et al. (1998). The KRR is very similar to SVM in the way that also uses kernels and combines them with ridge regression, but it is slower to test since it does not use support vectors to reduce complexity. When applied to our problem in terms of the Lagrange multipliers it can be formulated as the min-max optimization problem shown in equation 2.19.

$$\begin{aligned} &\underset{\eta}{\text{minimize}} \quad \underset{\alpha}{\text{maximize}} - \alpha^\top (K_\eta + \lambda I)\alpha + 2y^\top \alpha \\ &\text{where} \qquad\qquad \eta \in \ \{0 \preceq \eta \wedge ||\eta - \eta_0||_2 \leq \Lambda\} \end{aligned} \tag{2.19}$$

In equation 2.19, $K_\eta$ is the kernel originated from a combination $\eta$ of weights, $\eta_0$ and $\Lambda$ are two model parameters. According to Cortes et al. (2009) for any fixed $\eta$ the optimum is given by:

$$\alpha = (K_\eta + \lambda I)^{-1} y) \tag{2.20}$$

By plugging this equation to 2.19 we obtain:

$$\begin{aligned} &\underset{\eta}{\text{minimize}} \quad F(\eta) = y^\top (K_\eta + \lambda I)^{-1} y \\ &\text{where} \quad \eta \in \ \{0 \preceq \eta \wedge ||\eta - \eta_0||_2 \leq \Lambda\} \end{aligned} \tag{2.21}$$

This equation will allow us to find the weights by doing Algorithm 6.

1 `Algorithm(`*Projection-based Gradient Descent Algorithm*`);`
**input :** $\eta = \frac{1}{P}$, $K_n$,$n \in [1,P]$, $\upsilon \in [0,1]$;
2 **begin**
3  **while** $|\eta_{old} - \eta| > \epsilon$ **do**
4  $\eta_{old} = \eta$;
5  $\eta = -\upsilon * \nabla F(\eta) + \eta$;
6  $\eta = \eta * (\lambda / |\eta|)$ ;
7  **end**
8 **end**

**Algorithm 6:** Projection-based Gradient Descent Algorithm

$\epsilon$ is our threshold to stop the gradient descent and $\upsilon$ is the the step size. For solving this algorithm we must find the values for $\alpha$, this is done by using a solver for the SVM or CSSVM. We need the $\alpha$ because for any $k \in [1, p]$, the partial derivative of $F : \eta \to y^\top (K_\eta + \lambda I)^{-1} y$ with respect to $\eta_k$ is given by equation 2.22:

$$\frac{\partial F}{\partial \eta_i} = -2\alpha^\top (\sum_{r=1}^{P} (\eta_r K_r) * K_i)\alpha \tag{2.22}$$

This model consists in the creation of several kernels using the HTRBF, Gaussian and Anova, and apply them to each feature. Then we find the weight for each kernel created by solving 2.19 using 6. We then sum all kernels multiplied by the corresponding weight (2.17) to create our final kernel that can be used by the SVM equations (2.10) or (2.11).

## 2.7 Summary

In this Section we explained all the fundamental concepts that the reader needs to understand before reading the following Sections. We also talked about solutions to problems like class unbalancing. For the Sections that do not contain the full information on a subject, we always give a reference to the best paper or book we found on that area. In the next chapter we discuss about related work highlighting their focus and methods applied to breast image classification. We categorize them by area of interest.

# Chapter 3

# Related Work

In this chapter we will be discussing and listing works that are related to this work. We start by explaining why we did choose MKL and cite works in the area of classification. We will also discuss the human error and how relevant it is and list works that we can relate to ours. This chapter should be mainly used for finding the papers with the state of the art solutions for problems discussed in this work.

## 3.1 Initial Inspiration

Advances in technology have helped reduce tumors classification errors along the years. A number of research works have been able to discriminate between malignant and benign cases. Dhawan et al. (1996) uses automatically extracted features from mammograms and achieves an area under the ROC curve around 60%. Aarthi et al. (2011) uses a mixture of automatically extracted features and clinical features, achieving an accuracy of 86%. In the literature, some of the best results are achieved with SVM as is shown in Wei et al. (2005) and Ferreira et al. (2015) with ROC AUC of 83% and 85% respectively. Most work on SVMs use a single kernel, which may not be suitable for some data, because variable types can vary. In mammography data, features can be numeric, boolean, integer or categorical, depending on the way they are annotated. Recent work done by Augusto (2014) has shown that the use of MKL can help on the classification of mammography images, by employing a modified SVM, where, instead of using a single kernel, various kernels are used depending on the feature type (numerical, nominal or others). Augusto (2014) achieved 87% ROC AUC. The work done by Daemen et al. (2012) has also shown that MKL strategies can be useful to generate better models for clinical data.
In this work, like in the ones previously cited, we use SVM and MKL to do classification of ROI, and the data is also made of automatically extracted image features and heterogeneous clinical data manually annotated from the mammograms. But instead of focusing our work only in classification, we explore and study all the steps that are required to do classification of ROI in a mammogram. Because of that, this work combines several areas of research that are usually found separated. We take on the problem of unbalanced data with CSSVM and use a more

recent MKL algorithm, that according to Gonen and Alpayd (2011), is able to achieve better results than the one used by Augusto (2014).

## 3.2 Human Classification

Nowadays there is no scientific area that does not benefit from the utilization of computer systems, and obviously health is no exception, but due to the risk involved we do not expect that the CAD systems will be able to substitute and or do the work by the health specialists on the short term. We just want to build CAD systems that help these specialists do their work even if it just simplifies or creates better conditions to work. Another advantage of CAD systems is the inherent condition of being able to store everything that is being done. This data can be later worked on and internally each health institution may be able to detect misdiagnoses and more easily improve the performance of the mammography specialists, by telling them where and when were they wrong. One example of this is a system that automatically highlights the more relevant ROI, it does not have to be 100% accurate at detecting ROI with malignancy, because even if normal zones are highlighted the specialist can just ignore them and focus on the important ones. According to Barlow et al. (2004) a health specialist working with mammograms have a mean of True Negative rate of 90.1% and True Positive rate of 81.6%. Studies have shown that there is a 40% disparity among radiologist recall results and an even greater value of 45% for the False Positive (over treatment). Globally it is known that the ability of radiologists to detect cancer varies by as much as 11%. These values tell us that there is a large room for improvements and that CAD systems can be helpful decreasing these disparities, also the human average recall value of 81% is inferior to the one obtained by SVM 85% (Ferreira et al. (2015)).
To illustrate the human error and assuming each radiologist will see only 5 cancers in 1000 mammograms. With a recall of 81.6%, the radiologist will have a false negative rate of 1 per 1000 mammograms. If we assume that a health professional looks at least to 500 mammograms per year then every two years one person will go home without knowing that has cancer, and may die because of that mistake. Even worse is that there can be no feedback to the health professional about this mistake and because of that there is also no improving, repeating statistically the same error every two years.

## 3.3 The state of the art

We will be listing now the most recent works in the area of X-ray images and MKL. Most of the works we describe here use one or several of the following datasets; Mammographic Image Analysis Society (MIAS (2017)), Digital Database for Screening Mammography (Heath et al. (1998, 2000)), InBreast (Moreira et al. (2012)) and Breast Cancer Digital Repository (Moura et al. (2013); Ramos-Pollán et al. (2012); Moura and Guevara López (2013)). For each work,

we refer to the model or strategy used, dataset and objective. The metrics we will be using for comparison will be accuracy, AUC ROC or recall, different metrics are necessary because not all works use the same metric.

### 3.3.1 Detection and Labeling of Abnormalities

The following works are specialized in the detection or classification of a single type of finding like masses, calcification or structural distortion.

- Liu et al. (2016a) Presents an SVM model for recognition of architectural distortion in mammograms. The data contains 231 malignant masses from DDSM of those only 69 are architectural distortion. 60% were used for training and 40% for testing and obtaining up to 91,67% accuracy with SVM.
- Guo et al. (2016b) Created a model using 87 images from the MIAS Dataset. They were able to detect ROI in 81 of them by the enhance of abnormal mammograms using dual morphological top-hat operations with a non-flat structuring element, which is a method for image threshold, and a Neural Network for classification.
- Muthuvel et al. (2017) Created a micro calcification cluster detector using a multi scale products based Hessian matrix. The database consisted of 234 mammograms from both MIAS and DDSM combined, containing a total of 171 clusters of micro calcifications in which they show a total of 166 True Positives and 106 False Positives. Although they detected almost all of the clusters, they also report a percentage of 45% False Positive per image.
- Nithya and Santhi (2017) Built a model based in a Decision Tree classifier for mammogram density measure and classification. 180 mammograms were randomly chosen from the MIAS database, being 60 images of fatty, 60 images of glandular and 60 images of dense breasts. Three-fold cross-validation was used to evaluate the classifier. They correctly classified 98% of the mammograms into three density classes.

### 3.3.2 Classification of Regions of Interest

All these works propose classifiers for ROI. The methods change from paper to paper and almost all the state of the art data mining algorithms are used. The type of problem also changes between the papers, because some of them consider images with no benign ROI and try to classify them as normal.

- Liu et al. (2010) Created a model using Linear Discriminant Analysis and SVM for mass classification, where 309 images are used from DDSM, 142 benign and 167 malignant. They report 65% accuracy with SVM in what they consider a difficult dataset.

- Hiba et al. (2016) Presents a model using C4.5 Decision Tree and the k-Nearest-Neighbour algorithm, achieving 90% accuracy on the classification of different types of masses. The dataset contained 196 mammograms depicting a true mass and the rest 392 being normal mammograms without any mass.

- Audithan et al. (2017) Use different types of entropy measures and ensemble classification with three types of classifiers; k-Nearest-Neighbour, Bayes Network, and SVM for detection of malignancy in mammograms. Using the MIAS dataset they achieved results ranging from 62% to 89% accuracy.

- Isikli Esener et al. (2017) Created an ensemble method using Fisher's Linear Discriminant Analysis, Linear Discriminant Classifier, SVM, Logistic Linear Classifier, Decision tree, Random Forest, Naïve Bayes, and k-Nearest-Neighbour for Breast Cancer Diagnosis. The dataset IRMA from Deserno et al. (2011) with 233 ROI, was used with cross-validation, where 90% were used for training and 10% as the test, the results of classifiers that show the top three performances were combined achieving up to 93,52% accuracy.

- Milosevic et al. (2017) Took on the problem of three class mammograms classification (normal, benign, malignant). They presented a CAD system based on gray-level co-occurrence matrices using a SVM classifier, Naive Bayes and K-Nearest-Neighbour. These three methods were compared by doing cross-validation and obtained 65%, 51.6%, 38.1% accuracy, respectively.

- Qiu et al. (2017) Created a deep learning method for classifying between malignant and benign masses, using an image dataset involving 560 ROI extracted from digital mammograms. With a 4-fold cross-validation method they achieved a ROC AUC of 79%.

- Suhail et al. (2017) Show a tree based model for the classification of mammographic benign and malignant micro calcification clusters. The experiment was in a subset of 129 ROI from the DDSM database, 71 images were malignant, whereas 58 were benign. They report 66 True Positive results and 51 True Negative resulting in 91% accuracy.

- Venkatalakshmi and Janet (2017) Proposed a model with Pseudo Zernike Moments and SVM for the classification of ROI using MIAS dataset. They report accuracy values of up to 99%.

### 3.3.3 Multiple Kernel Learning

The following works are the most recent in the area of MKL applied to X-ray images.

- Espinoza (2016) Shows a model for detection of architectural distortion and characterization of masses by the use of MKL. For the mass description they obtained an average of 92% accuracy in the DDSM dataset and 94% accuracy for INBreast using 2 fold cross-validation. For the architectural distortion detection they got up to 89% accuracy in the DDSM (10 fold cross-validation) and 89% accuracy for MIAS using Leave one out.

- Cao et al. (2017) Present a MKL method for detection of Lung nodules in clinical thoracic CT scans. The database used in this work is the LIDC-IDRI made by Armato et al. (2011), containing 1012 cases. The experiments were done using 10-fold cross-validation and achieved a mean of 87% ROC AUC.

- Narváez et al. (2017) Presented an automatic BI-RADS characterization of breast masses contained in a ROI using MKL. The datasets used were the DDSM (980 ROI) and INBreast (216 ROI). They achieved sensitivity of 96.2% and a specificity of 93.1% on the mass detection. Also showed averaged sensitivity rates between 87.4% and 96.7% and specificity between 85.6% and 96.7%, on the shape, margin and density descriptions.

- Wani and Raza (2017) Used MKL method for classification of mammograms using 300 mammograms from the MIAS dataset. With 5 fold cross-validation they achieved 86% of what we assume to be accuracy since they do not refer that in the article.

### 3.3.4 Feature Selection

Feature Selection can be one of the best ways to maximize the performance of a model. We will now be discussing some of the different works that presented new feature selectors. All the works cited here are related with data mining of mammogram images and divided by learning task.

#### 3.3.4.1 ROI Classification

- Beura (2016) Present a correlation based filter, tested in MIAS and DDSM dataset using several data mining methods, achieving up to 98% accuracy for both datasets.

- Devisuganya and Suganthe (2016) Presents a Hybrid Shuffled frog-PSO algorithm for feature selection, the dataset from MIAS was used. With Decision Trees they show a recall of 92%.

- Kumar and Balakrishnan (2016) Did experiments using Symmetric Stochastic Neighbor Embedding for feature selection, the MIAS dataset is used. Experiments with SVM achieve results up to 90% accuracy.

- Liu et al. (2016b) Used a multitask learning method for feature selection in the DDSM dataset, and achieves 91.79% accuracy using a sparse representation based classification method.

- Galván-Tejada et al. (2017) Used the BCDR dataset for testing the generic algorithm from Galgo, a R software package for feature selection. Three algorithms were used for testing: Random Forest (93% ROC AUC), Nearest Centroid (93% ROC AUC) and K-Nearest-Neighbors (96% ROC AUC).

- Tamrakar and Ahuja (2017) Use Discrimination Potentiality to do feature selection. The data used is from DDSM and MIAS datasets, and they report 100% accuracy for the MIAS dataset and 97% for DDSM using an SVM.

#### 3.3.4.2 Others

- Guo et al. (2016a) Developed a new fuzzy-rough feature selection algorithm for Breast Density Classification, their experiments in MIAS dataset with Bayes network achieved up to 68% accuracy.
- Kowsalya and Priyaa (2016b) Created a model for detection of Bilateral Asymmetry in mammograms, using MIAS dataset, the feature selection is done using a Particle Swarm Optimization (79% recall), Ant Colony Optimization (82% recall) and Artificial Bee Colony Optimization (89% recall), the performance of each selector was tested using an Artificial Neural Network classifier.
- Tan et al. (2016) Developed a modified sequential floating forward selection for feature selection and experimented using their developed CAD system for detection of mammographic lesions, achieving up to 92% recall.

### 3.3.5 Feature Extraction

The list of extracted features and method for feature extraction used by us is described in Chapter 2.2.2. Many works use several more features or just different ones than us. This distinction is very important because the difference in the results can be because of the different dataset or the final set of features and not the model. The dataset with the mammograms is the same in some works, but the dataset with the objects that are going to be classified can be completely different. Databases like MIAS and DDSM only contain clinical features and mammogram images and because of that extra features must be extracted from ROI. We divided the works by groups of extracted features. This work does not focus in the selection of best group of features so we will not discuss which combination would be the optimal, but in future works we expect to explore that area.

- **Gray-level or Intensity descriptor**
  - Aarthi et al. (2011)
  - Nithya and Santhi (2017)
- **Shape, Size and Texture using Gray Level Co-occurance Matrix**
  - Wei et al. (2005)
  - Kowsalya and Priyaa (2016b)

  - Liu et al. (2016b)
  - Milosevic et al. (2017)
  - Suhail et al. (2017)

- **Intensity Descriptors And Texture Descriptors**
  - Guo et al. (2016b)
  - Cao et al. (2017)

- **Clinical Data**
  - Ferreira et al. (2015)

- **Intensity Descriptors , Texture Descriptors , Shape and Location, Clinical data**
  - Moura et al. (2013)
  - Moura and Guevara López (2013)
  - Augusto (2014)
  - **This work**

- **Gabor features and Texture Descriptors**
  - Liu et al. (2016a)
  - Tan et al. (2016)
  - Tamrakar and Ahuja (2017)

- **Wavelet Features and Texture Descriptors**
  - Beura (2016)
  - Kumar and Balakrishnan (2016)
  - Audithan et al. (2017)

- **Zernike-Wavelets and Gaussian Markov Random Field**
  - Devisuganya and Suganthe (2016)

- **Zernike-Wavelets**
  - Espinoza (2016)
  - Venkatalakshmi and Janet (2017)
  - Narváez et al. (2017)

- **Others**[1]
  - Trabelsi Ben Ameur et al. (2016)
  - Wani and Raza (2017)

[1] A special mention should be done to these works, because they use a combination of all the types of features that we refer in this paper.

## 3.4 Analysis

Our work is based on the previous results from Augusto (2014), we wanted to remake the experiment using a bigger dataset and a more recent MKL algorithm. We also do research and experiments in segmentation and feature extraction, this allows us to have a better perspective of the problem since we study and apply all the steps necessary to obtain classification of a ROI. Many works that we found do not use MKL and even less do a ROI classification, the most common type of work is specialized in the classification of one unique type of ROI, for example the work of Qiu et al. (2017) only uses masses. This type of strategy leads to very small datasets and may even lead to overfitted models. Our work is important because not only has experiments with different image databases but also does not specialize in any type of finding, because of that it is a much more generic and realistic model that can classify any type of finding in a mammogram.

Almost none of the cited works try to deal with the unbalanced characteristics of the data, we not only do that but also show results with a very unbalanced dataset. We also give more weight to the malignant class since a misclassification of a malignant object may lead to death. Kuusisto et al. (2015) studies this problem, using clinical data from mammograms to create a Naive Bayes network and estimate the probability of malignancy following a non-definitive breast core needle biopsy. They use a False Negative and False Positive weight of 150:1 and are able to increase Specificity while achieving 100% recall. Many papers say that the feature extraction is one of the most important parts of the experiment. The features we extracted are different from many other works, we think this happened because they are the ones present in the BCDR dataset, which is a dataset used in a very small number of works. Feature selection is another important part of our work and we noticed that almost no works use the same feature selector, almost all papers present a new solution for mammograms features selection.

According to the research done, the state of the art of mammogram classification is defined by an agglomeration of different works from different areas of research. This happens because of the several procedures that must be done to extract data from a mammogram, and ultimately use it for obtaining knowledge. Our work is one of the few that combines all that information into one single work, and suggests the creation of a system that alone is able to do all those procedures. Because of this, our work should be considered state of the art, since it combines state of the art algorithms to achieve one of the best solutions to the optimal classification of ROI in mammograms.

## 3.5 Summary

In this chapter we presented our position on the use of CAD systems in health care, and how they should be, we also presented the state of the art in X-ray image, we described the utility of our work on the present state of the art, and did recommendations on what can still be done and what paths should the research in this area follow. We expect that the reader is now able to

understand and criticize all our results and methods. In the next chapter we will describe the data we use and how the data-treatment was done to obtain the final objects. Results of our methods and algorithms are presented, and we finish the chapter with a comment on those while comparing them with other works but mainly to the work of Augusto (2014) since it is the work that better relates to this one.

# Chapter 4

# Weighted MKL applied to breast images

In this chapter, we will describe the methodology and data used in this work and how we apply the MKL method to our data. From now on to avoid any misinterpretations we will refer to positive +1 class as being malignant and negative -1 class as being the benign.

## 4.1 Data

As in other works we used two different datasets, Breast Cancer Digital Repository (BCDR) and Digital Database for Screening Mammography (DDSM).

### 4.1.1 BCDR

One of the datasets used for this experience was obtained by joining 4 different datasets from the Breast Cancer Digital Repository (BCDR (2017)).

> BCDR is a compilation of Breast Cancer anonymized patients' cases annotated by expert radiologists containing clinical data (detected anomalies, breast density, BIRADS classification, etc.), lesions outlines, and image-based features computed from Craniocaudal and Mediolateral oblique mammography image views.

The 4 datasets used from the BCDR repository were created in the works of: Moura et al. (2013); Ramos-Pollán et al. (2012); Moura and Guevara López (2013): bcdr_f02_features, bcdr_f01_features, bcdr_d01_features and bcdr_d02_features.

| I.V | Node | Calci | Micro | Archi | Stroma | Den |
|---|---|---|---|---|---|---|
| 1:177 | 0:452 | 0:432 | 0:606 | 0:779 | 0:726 | 1:165 |
| 2:221 | 1:356 | 1:376 | 1:202 | 1: 29 | 1: 82 | 2:222 |
| 3:187 | | | | | | 3:342 |
| 4:223 | | | | | | 4:79 |

Table 4.1: **I.V:** Image View, **Node:** is nodule, **Calci:** is calcification, **Micro:** is micro-calcification, **Archi:** is architectural distortion, **Stroma:** is stroma, **Den:** Density

| Feature | Malignant | Benign |
|---|---|---|
| is nodule | 164 | 192 |
| is calcification | 32 | 344 |
| is mic.calcification | 115 | 87 |
| is architectural distortion | 23 | 6 |
| is stroma | 67 | 15 |

Table 4.2: BCDR Lesion type malignant Count

The data from the combination of those datasets consists of a total of 904 unique objects with 36 features. The objects are several findings, from different patients. The 7 features described in table 4.1 and the Age are considered by BCDR to be clinical and general data. The remaining 29 are described in Figure 4.3 and were obtained from feature extraction from the regions of interest selected by radiologists. Details on the amount of malignant and benign cases for each lesion type can be seen in table 4.2. Before doing any data treatment objects with missing values were removed, maintaining 808 of our original 904 objects. This was done so that we did not have to train or test our model with missing values or predict them. In any case, the number of objects with missing values is very low. Our final dataset has 238 malignant cases and 570 benign.

### 4.1.2 DDSM

With the huge collection of mammogram images contained in the DDSM: Digital Database for Screening Mammography Heath et al. (1998, 2000) we created 2 datasets. The DDSM is organised by volumes of cases that contain images.

> Each volume is a collection of cases of the corresponding type. A case consists of between 6 and 10 files. These are an "ics" file, an overview "16-bit PGM" file, four image files that are compressed with lossless JPEG encoding and zero to four overlay files. Cancer cases are formed from screening exams in which at least one pathology proven cancer was found. benign cases are formed from screening exams in which something suspicious was found, but was determined to not be malignant (by pathology, ultrasound or some other means).

Table 4.3: BCDR Continuous Features

| | Age | Mean | Std_Dev | Max | Min |
|---|---|---|---|---|---|
| Min | 23 | 0.11 | 0.020 | 0.17 | 0 |
| Mean | 1.73 | 0.66 | 0.108 | 0.90 | 0.34 |
| Max | 2.94 | 0.99 | 0.294 | 1.00 | 0.81 |
| | Area | Perim | x_Centroid | y_Centroid | Circularity |
| Min | 129 | 40 | 0.019 | 0.080 | 0.03 |
| Mean | 25465 | 511 | 0.48 | 0.497 | 0.66 |
| Max | 829617 | 4039 | 0.99 | 0.911 | 1.05 |
| | Form | Solidity | Extent | Energy | Contrast |
| Min | 0.0003 | 0.1779 | 0.039 | 0.0033 | 0.4 |
| Mean | 0.0083 | 0.8772 | 0.621 | 0.0916 | 14 |
| Max | 0.0386 | 1.0000 | 0.859 | 0.8946 | 138 |
| | Variance | Homogeneity | Sum Average | Sum Variance | Sum Entropy |
| Min | 22 | 0.20 | 9 | 60 | 0.36 |
| Mean | 523 | 0.50 | 43 | 1858 | 2.68 |
| Max | 1014 | 0.95 | 63 | 3995 | 4.01 |
| | Skewness | Info.Correlation 2 | Info.Correlation | Diff Variance | Entropy |
| Min | -5.69 | 0.060 | -0.6417 | 0.4 | 0.39 |
| Mean | -0.33 | 0.61 | -0.1438 | 14 | 3.96 |
| Max | 6.11 | 0.96 | -0.0087 | 138 | 6.22 |
| | Correlation | Elongation | Kurtosis | Difference Entropy | |
| Min | -0.52 | 0.05 | -1.6 | 0.36 | |
| Mean | 0.38 | 0.68 | 1.6 | 56 | |
| Max | 0.96 | 0.98 | 52.3 | 89 | |

* For more information on the equations used for extracting the variables please refer to 2.2.2.

We Selected the benign and malignant volumes that use the scanner Lumisys and Howtek, we chose these because they were the ones that contained more cases and since the images were scanned by different machines, we decided that 2 datasets could be made.

#### 4.1.2.1 Terminology

One of the main reasons for using DDSM is that each image contains clinical features that we wanted to use in our experiments. we will now describe these features according to DDSM (2017).

**Subtlety** The subtlety value for a lesion may indicate how difficult it is to find the lesion, the bigger the easier (1 is "subtle" and 5 is "obvious").

**Mass Shape** :

**Round** A mass that is circular in shape.

**Oval** A mass that is elliptical.

**Lobulated** A mass that has contours with undulations.

**Irregular** The lesion's shape cannot be characterized.

**Distortion** Equivalent ot an architectural distortion in BI-RADS.

**Mass Margin** :

**Circumscribed** The margins are sharply demarcated with an abrupt transition between the lesion and the surrounding tissue.

**Ill Defined** Poor definition of the margins.

**Obuscured** One which is hidden by adjacent normal tissue.

**Spiculated** The mass is characterized by lines radiating from the margins.

**Microbulated** The margins contain small undulations.

**Calcification Type** :

**Punctate** These are circular, less than 0.5mm with well defined margins.

**Amorphous** Calcifications so small that cannot be characterized.

**Pleomorphic** Bigger than amorphous.

**Lucent Center** These are calcifications that range from under 1 mm to over a centimetre or more.

**Fine Linear Branching** These are thin, irregular calcifications, they also are discontinuous.

**Calcification Distribution** :

**Clustered** Used when multiple calcifications occupy a small volume of tissue.

**Linear** Calcifications arrayed in a line.

**Regional** These are calcifications scattered in a large volume of breast tissue.

**Diffuse** These are calcifications that are distributed randomly throughout the breast.

In tables 4.4 and 4.5 NA represents values that were not annotated.

#### 4.1.2.2 Howtek

The data from the combination of the Howtek volumes consists of a total of 1380 objects with 27 features where 717 objects are benign and 663 are malignant. Basic information about the clinical features can be seen in Table 4.4 and the automatically extracted features are in Table 4.6.

#### 4.1.2.3 Lumisys

The data from the combination of the Lumisys volumes consists of a total of 1372 objects with 27 features where 703 objects are benign and 669 are malignant. Information about the clinical features can be seen in Table 4.5, the automatically extracted features are in Table 4.6.

| | Class | |
|---|---|---|
| Shape | -1 | +1 |
| Distortion | 13 | 46 |
| Irregular | 77 | 289 |
| Lobulated | 141 | 45 |
| Oval | 145 | 43 |
| Round | 42 | 11 |
| Other | 3 | 8 |
| NA | 226 | 291 |

| | Class | |
|---|---|---|
| Margins | -1 | +1 |
| Circumscribed | 92 | 16 |
| Ill Defined | 144 | 175 |
| Obscured | 145 | 25 |
| Spiculated | 24 | 205 |
| Others | 21 | 10 |
| NA | 291 | 232 |

| | Class | |
|---|---|---|
| Subtlety | -1 | +1 |
| 1 | 74 | 71 |
| 2 | 167 | 152 |
| 3 | 210 | 153 |
| 4 | 177 | 159 |
| 5 | 89 | 128 |

| Distribution | +1 | -1 |
|---|---|---|
| Clustered | 240 | 168 |
| Linear | 15 | 43 |
| Segmental | 33 | 38 |
| Other | 6 | 1 |
| NA | 423 | 413 |

| Type | +1 | -1 |
|---|---|---|
| Amorphous | 17 | 50 |
| Branching | 33 | 2 |
| Pleomorphic | 175 | 189 |
| Punctate | 17 | 35 |
| Other | 8 | 11 |
| NA | 413 | 423 |

| Lesion | +1 | -1 |
|---|---|---|
| Calcification | 294 | 249 |
| Mass | 423 | 414 |

Table 4.4: Howtek Clinical Features

| | Class | |
|---|---|---|
| Distribution | +1 | -1 |
| Clustered | 235 | 116 |
| Linear | 11 | 2 |
| Segmental | 31 | 27 |
| Cluster-Linear | 8 | 10 |
| Regional | 5 | 10 |
| Other | 0 | 7 |
| NA | 413 | 496 |

| | Class | |
|---|---|---|
| Type | +1 | -1 |
| Amorphous | 53 | 20 |
| Branching | 14 | 16 |
| Pleomorphic | 150 | 91 |
| Lucent-Centered | 24 | 0 |
| Branching-Pleomorph | 3 | 19 |
| Punctate | 21 | 5 |
| Other | 33 | 11 |
| NA | 365 | 498 |

| | Class | |
|---|---|---|
| Lesion | +1 | -1 |
| Calcification | 339 | 171 |
| Mass | 364 | 498 |

| shape | +1 | -1 |
|---|---|---|
| Irregular | 28 | 201 |
| Oval | 140 | 75 |
| Lobulated | 100 | 86 |
| Round | 59 | 31 |
| Distortion | 16 | 43 |
| IrregularDistortion | 3 | 46 |
| Other | 16 | 14 |
| NA | 341 | 173 |

| | Class | |
|---|---|---|
| Margins | +1 | -1 |
| Circumscribed | 209 | 31 |
| Circum.-Obscured | 23 | 2 |
| Ill-Defined | 41 | 117 |
| Microbulated | 19 | 66 |
| Obscured | 24 | 8 |
| Ill-Spiculated | 5 | 22 |
| Spiculated | 10 | 186 |
| Ill-Obscured | 2 | 24 |
| Other | 8 | 19 |
| NA | 350 | 179 |

| | Class | |
|---|---|---|
| Subtlety | +1 | -1 |
| 1 | 11 | 22 |
| 2 | 77 | 67 |
| 3 | 208 | 151 |
| 4 | 172 | 130 |
| 5 | 235 | 299 |

Table 4.5: Lumisys Clinical Features

Table 4.6: Howtek and Luminsys Continuous Features

| | Howtek Features | | | | Lumisys Features | | | |
|---|---|---|---|---|---|---|---|---|
| Statistic | Mean | St. Dev. | Min | Max | Mean | St. Dev. | Min | Max |
| mean | 111.859 | 26.283 | 49.248 | 175.164 | 110.108 | 27.417 | 42.758 | 176.921 |
| std_dev | 14.792 | 6.819 | 3.478 | 38.278 | 17.622 | 9.145 | 3.576 | 56.971 |
| mode | 113.783 | 29.313 | 43.626 | 186.794 | 114.104 | 33.125 | 32.416 | 193.416 |
| median | 112.132 | 26.803 | 48.179 | 177.214 | 110.573 | 28.636 | 41.237 | 179.479 |
| maximum | 161.357 | 30.817 | 69.685 | 239.724 | 167.824 | 31.389 | 81.809 | 247.603 |
| minimum | 72.941 | 27.145 | 3.790 | 137.253 | 64.100 | 28.785 | 0.000 | 136.420 |
| kurtosis | −0.204 | 0.713 | −1.446 | 2.490 | 0.325 | 2.632 | −1.541 | 33.557 |
| skewness | 0.058 | 0.526 | −1.612 | 1.600 | 0.138 | 0.773 | −2.138 | 5.167 |
| area | 186,457 | 176,621 | 3,309 | 923,070 | 198,443 | 217,232 | 944 | 1,409 |
| perimeter | 1,6102 | 820.749 | 225.602 | 3,951 | 1,587 | 875 | 138.400 | 4,468 |
| x_centroid | 0.514 | 0.222 | 0.027 | 0.970 | 0.498 | 0.269 | 0.021 | 0.982 |
| y_centroid | 0.524 | 0.146 | 0.135 | 0.916 | 0.488 | 0.148 | 0.089 | 0.919 |
| circularity | 0.717 | 0.093 | 0.204 | 0.911 | 0.802 | 0.179 | 0.076 | 0.982 |
| elongation | 0.740 | 0.152 | 0.195 | 0.993 | 0.764 | 0.147 | 0.110 | 0.992 |
| form | 0.001 | 0.001 | 0.0003 | 0.008 | 0.001 | 0.001 | 0.0002 | 0.015 |
| solidity | 0.958 | 0.040 | 0.628 | 0.994 | 0.947 | 0.078 | 0.332 | 0.995 |
| entropy | 5.779 | 0.657 | 3.720 | 7.400 | 5.931 | 0.702 | 3.580 | 7.501 |
| contrast | 0.033 | 0.011 | 0.0005 | 0.108 | 0.059 | 0.020 | 0.007 | 0.180 |
| correlation | 0.930 | 0.062 | 0.386 | 0.995 | 0.894 | 0.094 | 0.390 | 0.994 |
| energy | 0.496 | 0.176 | 0.175 | 0.998 | 0.443 | 0.179 | 0.155 | 0.975 |
| homogeneity | 0.983 | 0.006 | 0.946 | 1.000 | 0.970 | 0.010 | 0.910 | 0.997 |

* For more information on the equations used for extracting the variables please refer to 2.2.2.

### 4.1.3 Data Treatment

Standardization was applied to the data followed by winsorization, a technique to reduce the impact of outliers and turn them into usable values (Ghosh and Vogt (2012)). This is possible because after standardization, data is centered in zero and reflect how far the data is from the average (the Z-Score). In our case the maximum allowed Z-Score is $\pm 3$, which only affects 2% of data according to a normal distribution. In this case, after winsorization, the value 3.6 would be transformed into 3.

### 4.1.4 Methodology

To test the performance of our models the methodology shown in algorithm 7 was applied. With this methodology we make sure that no data from a test set is used in any part of the tuning and training processes, per fold. During the training we used the solvers from the SVM implementation of Chang and Lin (2011) and Cost-Sensitive SVM implementation of Masnadi-Shirazi et al. (2012). These solver were used on Non-Linear MKL created by 2.19 and implemented based on the work of Gonen and Alpayd (2011). Our implementation allowed us to change the SVM solver according to the type of problem. The weighted version was only applied to BCDR because the classes for these datasets are unbalanced.

```
1  Algorithm(MKL Methodology);
2  begin
3      Choose number of Folds N;
4      Create N Folds with a train and test set;
5      for each Fold do
6          Apply Feature Selection on the Training set;
7          Create x subfolds from training set;
8          for each combination of Parameters do
9              for each Subfold do
10                 Train on the training Subfolds;
11                 Test on the test Subfold;
12             end
13         end
14         Select the combinaton of Parameters that produced best results across all subfolds;
15         Train Model with the Parameters;
16         Test on Test Set;
17         Store True Positives, True Negatives, False Negatives, False Positives;
18     end
19 end
```

**Algorithm 7:** MKL Methodology

The parameters we refer to in algorithm 7 are the following.

- $C$ From the SVM Equation 2.8.
- $\epsilon$ From MKL Equation 6.
- The number of features to be removed from the dataset.
- $Weights$ for margin and error rate values, when training with the CSSVM solver equation 2.9.

$C$, $\epsilon$ and $Weights$ tuning was controlled by arrays of possible values, and the tuning of features was controlled by the condition: if the results do not improve after removing a feature stop removing features.
For the experiment we used five folds for the BCDR datasets, each fold size can be seen in Table 4.7. For the DDSM datasets we opted for the ten folds, the sizes of each fold can be found in Tables 4.8 and 4.9. The difference in the number of folds between the BCDR and DDSM is due to the reduced number of malignant objects in BCDR, if we had done 10 fold we would end up with a very small amount of malignant objects per fold. This way all the folds across the 3 datasets have approximately 30 objects on the test set and 100 on the training set. Also the number of malignant objects per test set in both datasets ranges from 16 to 10 objects.

| Fold | Set Type | Benign Objects | Malignant Objects |
|---|---|---|---|
| 1 | Train | 96 | 23 |
| 1 | Test | 22 | 10 |
| 2 | Train | 102 | 26 |
| 2 | Test | 20 | 12 |
| 3 | Train | 80 | 48 |
| 3 | Test | 18 | 14 |
| 4 | Train | 89 | 39 |
| 4 | Test | 22 | 10 |
| 5 | Train | 96 | 32 |
| 5 | Test | 21 | 11 |

Table 4.7: BCDR Fold Details

| Fold | Set Type | Benign | Malignant |
|---|---|---|---|
| 1 | Train | 53 | 55 |
| 1 | Test | 13 | 14 |
| 2 | Train | 61 | 47 |
| 2 | Test | 11 | 16 |
| 3 | Train | 54 | 54 |
| 3 | Test | 14 | 13 |
| 4 | Train | 55 | 53 |
| 4 | Test | 16 | 11 |
| 5 | Train | 52 | 56 |
| 5 | Test | 12 | 15 |
| 6 | Train | 59 | 49 |
| 6 | Test | 14 | 13 |
| 7 | Train | 55 | 55 |
| 7 | Test | 16 | 11 |
| 8 | Train | 55 | 53 |
| 8 | Test | 14 | 13 |
| 9 | Train | 64 | 44 |
| 9 | Test | 12 | 15 |
| 10 | Train | 54 | 54 |
| 10 | Test | 12 | 15 |

Table 4.8: DDSM Howtek objects per fold

| Fold | Set Type | Benign | Malignant |
|---|---|---|---|
| 1 | Train | 60 | 48 |
| 1 | Test | 13 | 14 |
| 2 | Train | 54 | 54 |
| 2 | Test | 11 | 16 |
| 3 | Train | 59 | 49 |
| 3 | Test | 14 | 13 |
| 4 | Train | 62 | 46 |
| 4 | Test | 13 | 14 |
| 5 | Train | 55 | 53 |
| 5 | Test | 11 | 16 |
| 6 | Train | 60 | 48 |
| 6 | Test | 12 | 15 |
| 7 | Train | 53 | 55 |
| 7 | Test | 14 | 13 |
| 8 | Train | 52 | 56 |
| 8 | Test | 13 | 14 |
| 9 | Train | 53 | 55 |
| 9 | Test | 15 | 12 |
| 10 | Train | 57 | 51 |
| 10 | Test | 14 | 13 |

Table 4.9: DDSM Lumisys objects per fold

## 4.2 Feature Selection

For the feature selection (line 4 of algorithm 7), we ranked features combining different strategies. Since all give a different output, we have less chances of overfiting our model since we take into consideration all the outputs. Overfiting the feature selection decreases our performance since the test set is never used while doing the feature ranking. We use the UFilter, the Recursive Feature Elimination from the R caret package (Caret) and feature-feature and feature-class correlations. The final rank of each feature is the sum of the positions achieved in each applied strategy (the lower the better). This process is describe in Algorithm 8.

```
1  Algorithm(Feature Selection Algorithm);
   input : FeaturesList, FeatureSelectionMethods;
2  begin
3      for each Method in FeatureSelectionMethods do
4          FeaturesRank.add(Rank(Method,FeaturesList));
5      end
6      for each Feature in FeaturesList do
7          tempFeatureRank = SumRank(Feature in all FeaturesRank));
8          RankResults.AddFeature(Feature, tempFeatureRank);
9      end
10     return Sort(RankResults);
11 end
```

**Algorithm 8:** Feature Selection Algorithm

.

During the tuning we remove up to 20 of the least interesting features and calculate which number showed best results. The mean values of features remaining per fold around all folds and datasets was of 19, being 17 the lowest value of remaining features.

| Rank | BCDR | DDSM-Lumisys | DDSM-Howtek |
|---|---|---|---|
| 20 | Minimum | Shape | Solidity |
| 19 | Is Microcalcification | Form | Subtlety |
| 18 | Entropy | Subtlety | Lesion |
| 17 | Sum Variance | Type | Type |
| 16 | Mean | Energy | Shape |
| 15 | Breast density | Circularity | Energy |
| 14 | Is architectural distortion | Mean | Standard deviation |
| 13 | Sum Average | Kurtosis | Form |
| 12 | Kurtosis | Solidity | Y Centroid |
| 11 | Difference Entropy | Elongation | Perimeter |
| 10 | Variance | Standard deviation | Contrast |
| 9 | Homogeneity | Perimeter | Mean |
| 8 | Image View | Skewness | Homogeneity |
| 7 | Energy | Median | X Centroid |
| 6 | Skewness | Minimum | Kurtosis |
| 5 | Age | Homogeneity | Elongation |
| 4 | Standard deviation | Contrast | Skewness |
| 3 | Y Centroid | Mode | Median |
| 2 | X Centroid | X Centroid | Minimum |
| 1 | Elongation | Y Centroid | Mode |

Table 4.10: Features Rank

Table 4.10 contains the 20 most important features of each dataset, it was built by sorting all the features by the average rank position in each fold. To better understand the results we use colors to differentiate some features. White features are the ones that got similar ranks across all the datasets, blue ones got similar results across 2 datasets and yellow are the features that appear in only one dataset.
The differences in the ranking of the features may be due to several causes:

- The unbalance of the classes in BCDR.
- Different nature of the images between DDSM and BCDR.
- Different Scanner used in both datasets of DDSM.
- There may be differences in the method used for feature extraction since we do not have access to the source used for the extraction of the BCDR features.
- BCDR contains features like age that are not in DDSM.

Since the number of remaining features for each fold was much smaller than the original (BCDR: Original 36; DDSM: Original 27) and got improvement during the tuning phase we can conclude that our feature selector is helping in the process of classification.

# Chapter 5

# Results

This section will be divided in three parts each one specific for a dataset. For each dataset we will discuss the results in Table 5.2, followed by the ROC AUC curves and PR curves that can be found in the next pages for better visualization. We will also comment on the statistical difference of each result according to the the P-Value, these results can be found in Table 5.1.

| | BCDR | Lumisys | Howtek |
|---|---|---|---|
| SVM MKL | 0.3 | $3e^{-2}$ | $2.8e^{-6}$ |
| SVM CS-MKL | $1.1e^{-5}$ | NA | NA |
| MKL CS-MKL | $4.6e^{-3}$ | NA | NA |

Table 5.1: P-Values

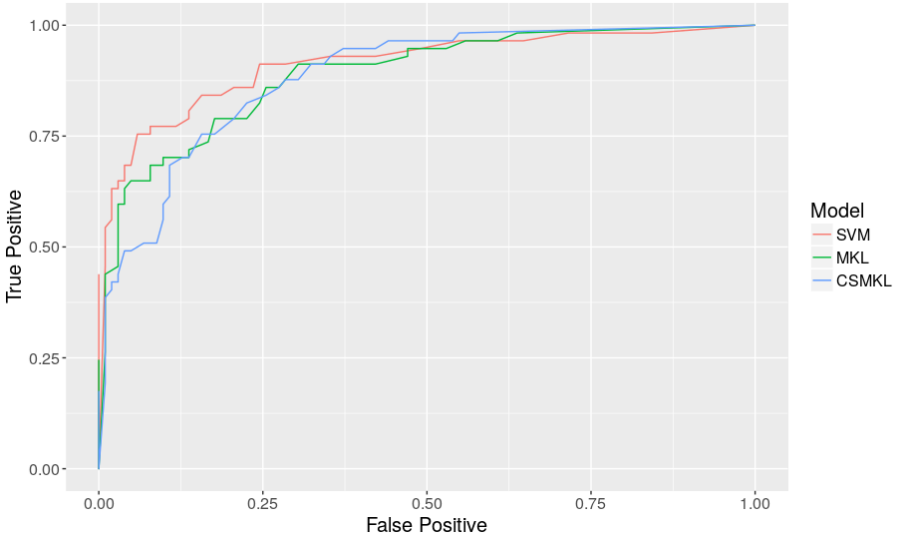
| BCDR Results | Accuracy | AUC | FOR | Precision | Specificity | Recall |
|---|---|---|---|---|---|---|
| SVM | 0.83 | 0.90 | 0.80 | 0.94 | 0.98 | 0.57 |
| MKL | 0.82 | 0.88 | 0.82 | 0.83 | 0.93 | 0.63 |
| CSMKL | 0.78 | 0.88 | 0.87 | 0.67 | 0.78 | 0.78 |

| DDSM Howtek | Accuracy | AUC | FOR | Precision | Specificity | Recall |
|---|---|---|---|---|---|---|
| SVM | 0.57 | 0.59 | 0.55 | 0.63 | 0.76 | 0.38 |
| MKL | 0.62 | 0.67 | 0.62 | 0.62 | 0.62 | 0.62 |

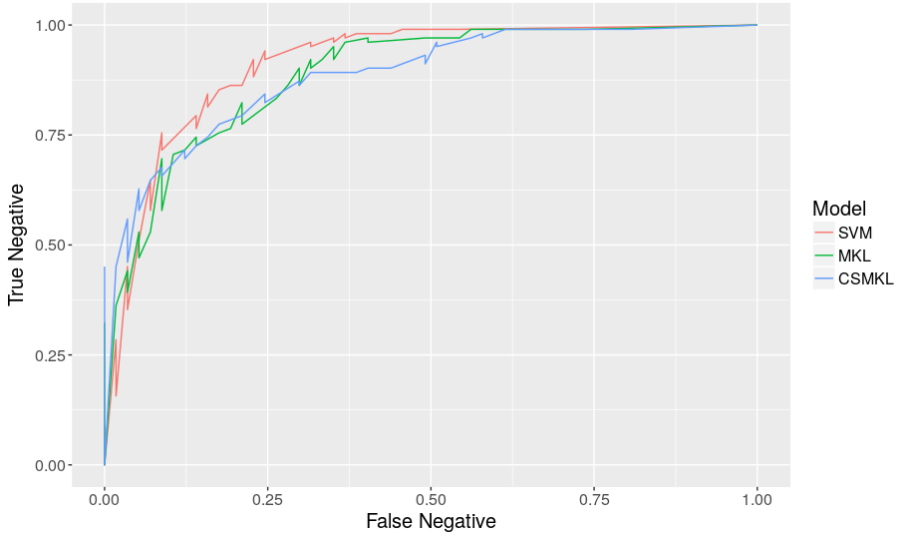
| DDSM Lumisys | Accuracy | AUC | FOR | Precision | Specificity | Recall |
|---|---|---|---|---|---|---|
| SVM | 0.58 | 0.62 | 0.56 | 0.62 | 0.69 | 0.5 |
| MKL | 0.78 | 0.83 | 0.77 | 0.80 | 0.78 | 0.78 |

Table 5.2: Experimental Results

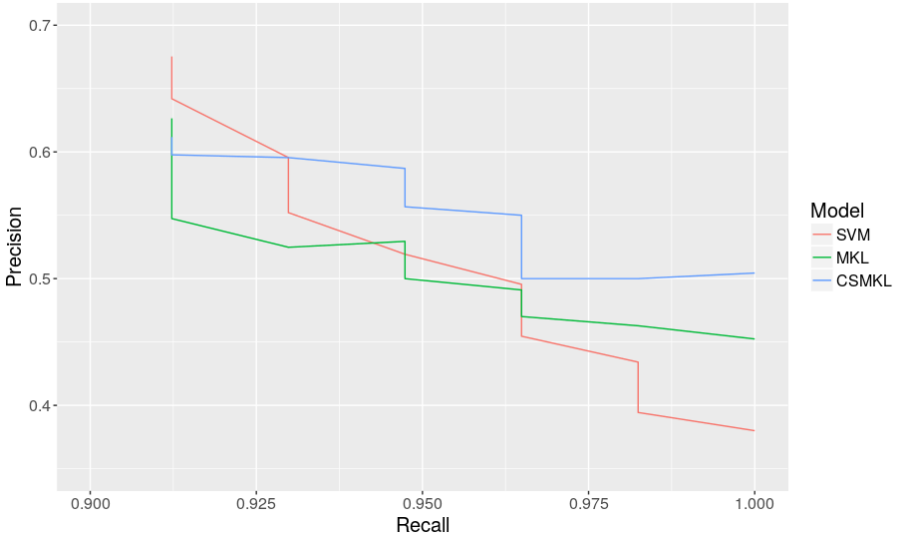
Figure 5.1: BCDR Dataset Curves



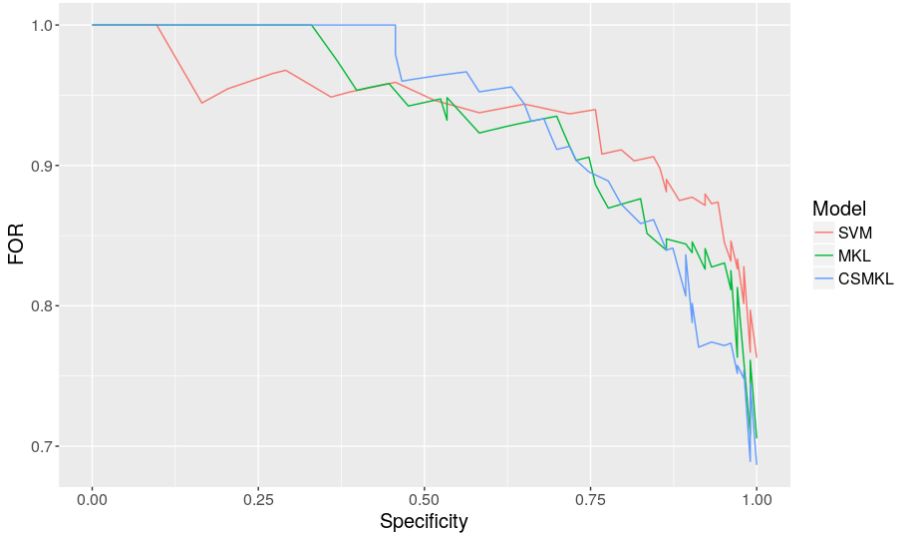

(a) ROC (Malignant)




(b) ROC (Benign)

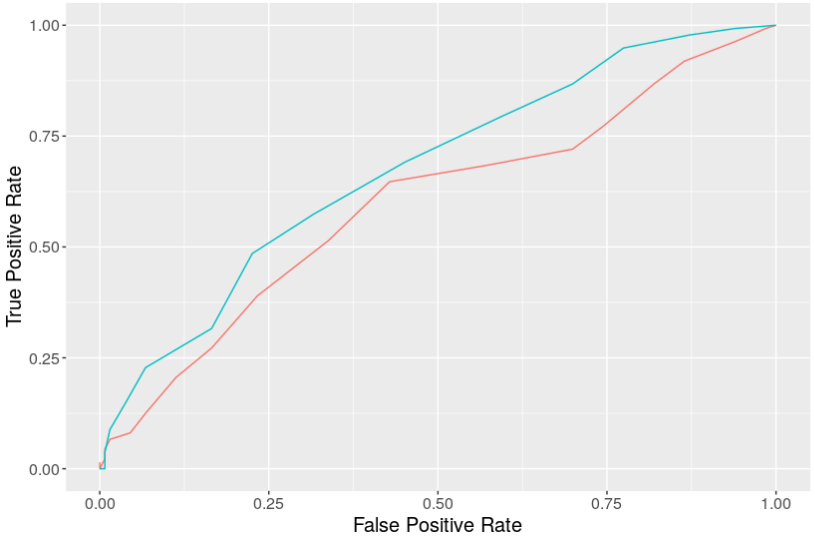



(c) Precision Recall

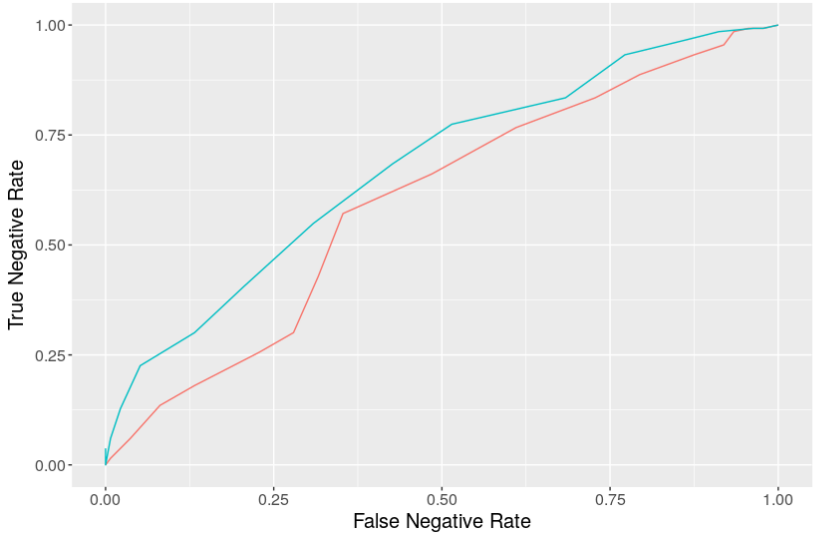



(d) FOR Specificity

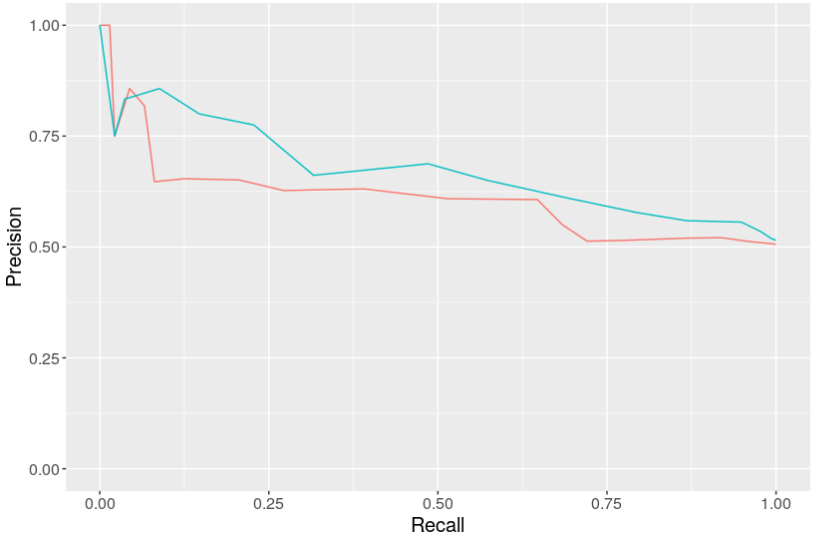
Figure 5.2: DDSM Howtek Dataset Curves



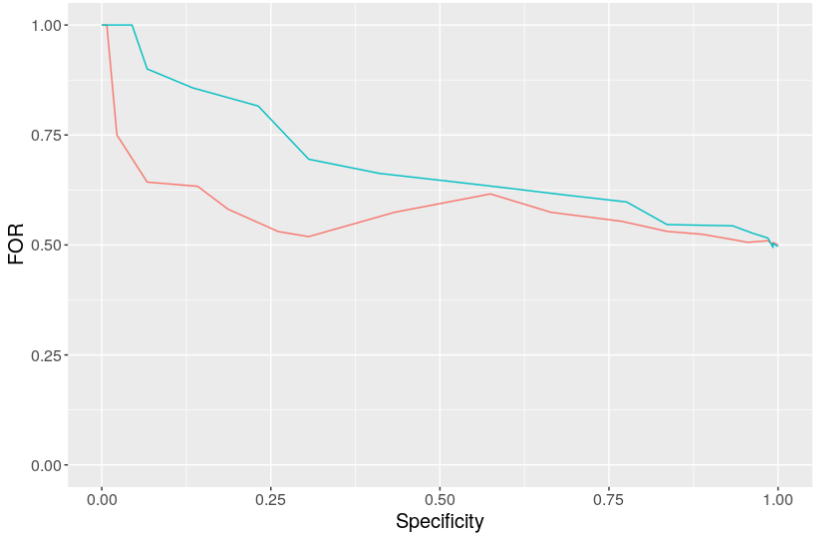

(a) ROC (Malignant)




(b) ROC (Benign)

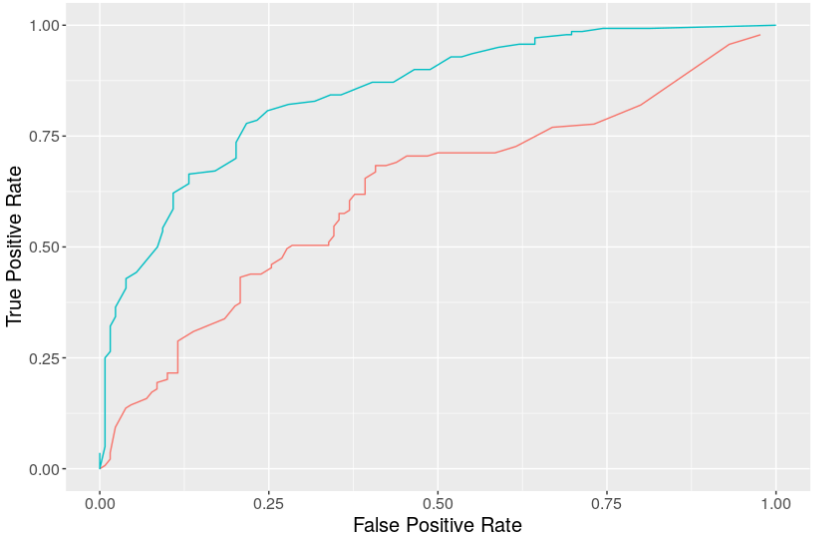



(c) Precision Recall

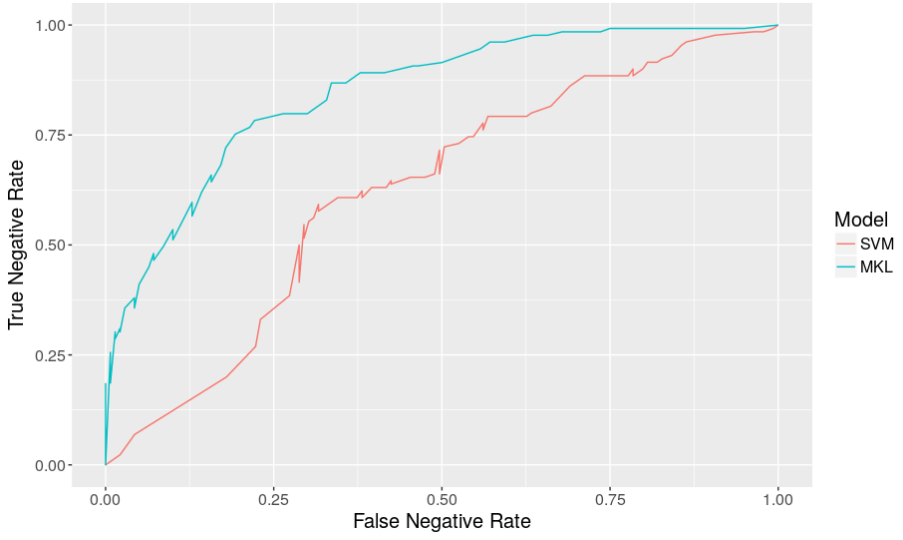



(d) FOR Specificity

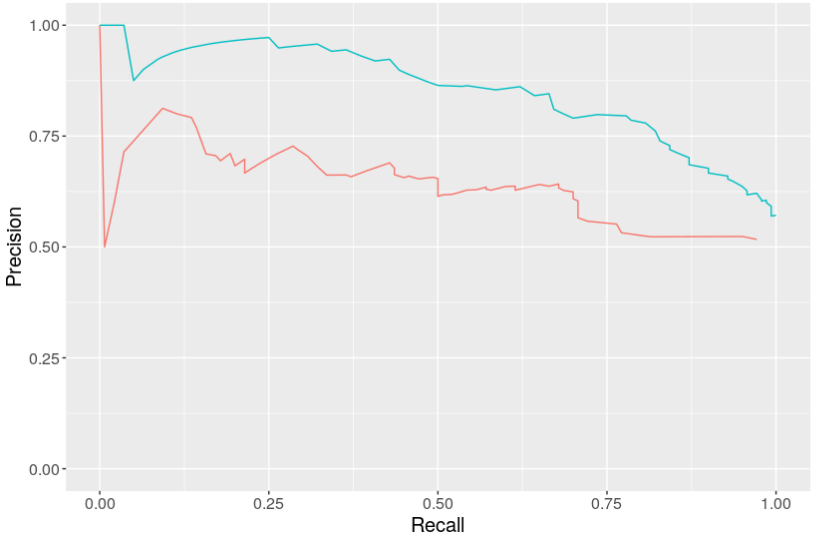
Figure 5.3: DDSM Lumisys Dataset Curves



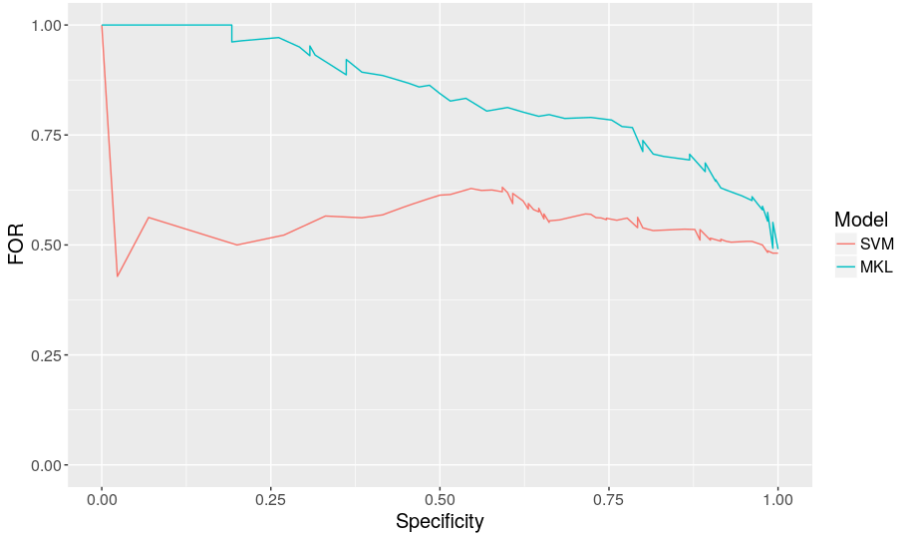

(a) ROC (Malignant)




(b) ROC (Benign)




(c) Precision Recall




(d) FOR Specificity

## 5.1 BCDR

Our objective in this experiment was to obtain the best possible recall for the malignant class that is the minority in the BCDR dataset. According to He and Garcia (2009) increasing recall of the results is usually followed by a decrease in the precision, because the model must allow more Negative objects to be misclassified. This becomes problematic when the number of negatives (like in this dataset) is higher than the number of positives, because the number of False Positives will grow faster than the number of True Positives. That can be seen in Figure 5.4 where two different hiperplanes divide the data. The normal (black) hiperplane divides the data while keeping an equal number of misclassified objects for both classes, the hiperplane with weights (green) gives more weight to the minority but greatly increases the number of objects that will be misclassified.

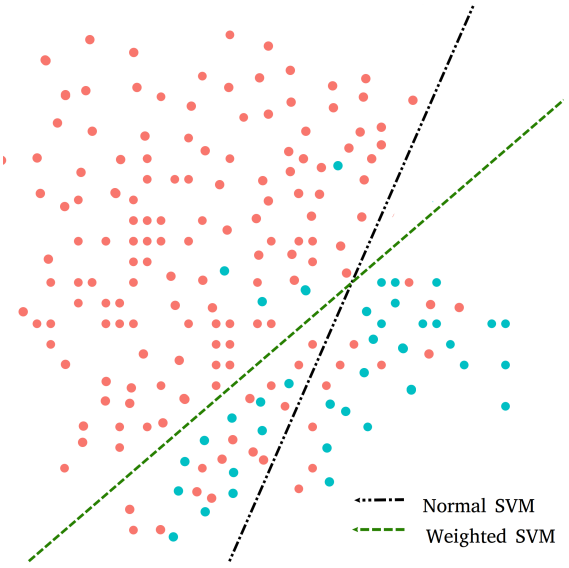



Figure 5.4: Hypothetical example of hyperplanes dividing unbalanced data.

### 5.1.1 Table Results

The values in Table 5.2 were calculated using the results from the folds of each dataset at 0.5 threshold. The results shown for the BCDR experiment in Table 5.2 suffer from the same problem that is described in Chapter 2.4.3 (Unbalanced Data Metrics), this can be easily seen in the difference between the specificity and recall for the SVM. Accuracy and AUC show that SVM has a general better performance, but if we focus in the recall and specificity we see that CSMKL was the model that was able to achieve a better result in the classification of the malignant class, achieving 21% more recall than SVM. These results also shows that to increase the recall we may lose precision but it was almost a symmetrical loss because CSMKL lost 27% of precision when compared to SVM (only 6% more than the recall gain).

### 5.1.2 Curves Results

It is shown in the (Positive)ROC curve of Figure 5.1a, that CSMKL is able to achieve higher rates (0.9) of true positives before the SVM, and in the (negative)ROC curve of Figure 5.1a MKL is better for lower rates (0.1) of False Negatives. Obviously this means that SVM has better results when there is a high specificity for the benign class, but CSMKL or MKL yield

better results at higher recall rates for the malignant class. We zoomed in the precision recall curve in Figure 5.1c to better show the difference between the 3 models when near 100% recall for the malignant class is required, and prove that CSMKL is the better of the three models for this type of problem, the difference in the precision between SVM and other Models near 100% recall is up to 15%. Comparison between CSMKL and SVM has $p < 0.01$, the same is true when comparing CSMKL with MKL. SVM and MKL obtain statistically similar results, but the results for CSMKL show that the model is different from any other and its use will drastically change the obtained results. This proves that for unbalanced datasets CSMKL can achieve better results by adjusting the weights according to the metrics we want to minimize or maximize.

## 5.2 DDSM

For the DDSM datasets we did not use CSMKL because both the datasets were balanced and the use of weights was achieving 100% benign|malignant recall and 0% recall for the other class. MKL got better results than SVM in both datasets. We should take into account that the Howtek and Lumisys are both datasets that are difficult to classify, SVM only got $0.58 \pm 1$ accuracy which means that the results can be considered random.

### 5.2.1 Howtek

When comparing the results from SVM and MKL we can see that SVM has beaten MKL in all the metrics that focus on the correct classification of the benign Class (precision and specificity), on the other hand MKL achieves better accuracy and AUC while having more than 0.2 recall than the SVM. (0.38 SVM to 0.62 MKL). The curves that can be found in Figure 5.2 show that MKL is always superior to the SVM, also the P-Value of MKL and SVM $p < 0.01$ tells us that the results are indeed statistically different and that there is no denial according to these results that MKL does show better performance than SVM when dealing with balanced data.

### 5.2.2 Lumisys

The results from this experiment are the best we have, the MKL was able to beat SVM in all the metrics we show in Table 5.2. The difference in the accuracy between both models is of 20% and not only MKL is able to increase the recall without losing any specificity. Like in the Howtek experiment the curves in Figure 5 show that MKL is superior in all the moments of the curve and that can also be seen in the AUC value of SVM (0.62) and the AUC of MKL (0.83). Also like in the experiments before it achieved a $p < 0.01$ which means that the results are without doubt different from the ones of SVM.

## 5.3 Analysis

We started this work with the objective of improving the results achieved by Augusto (2014) and we were able to do that not only with the MKL but also with the SVM, in a unbalanced dataset. Our work with feature selection also achieved very good results and the ranking system was able to improve the results of our MKL and SVM models. Then we took on the challenge of dealing with unbalanced datasets, the results for that experience were somewhat bellow what we were expecting, this was mainly because of the difficulties that come from experimenting and explaining results of an unbalanced dataset. Finally in both experiments with the DDSM dataset we were able to decisively achieve better results with MKL than with an SVM.
By looking at the results shown in 5, we concluded that MKL and CSMKL must be used in different types of datasets. CSMKL got better results when working with unbalanced data and was able to minimize the number of False Negatives to a value near 0 while maintaining an acceptable number of False Positives 50%. We now have a model that can correctly classify all the malignant cases while only misclassifying half of the benign cases. We are also able to prove that our MKL model can achieve significantly better results than a simple SVM in both of the DDSM datasets.
We achieved better results than other ones on the DDSM dataset and used a higher amount of objects for our experiments. Suhail et al. (2017) achieve 91% accuracy but uses only 129 ROI from the DDSM database and Liu et al. (2010) use 309 images from DDSM and achieve 65% accuracy.

## 5.4 Summary

In this chapter we presented all the work done by us, from the description of the data to the methodology and results, did some comments on which objectives were achieved and where we think that we failed, while comparing our results and methods to other works. The results we obtained where very satisfactory because they allowed us too confirm both of our initial thoughts that MKL was an improvement of SVM and that CSMKL was able to better classify unbalanced data. In the next chapter we will focus on discussing the areas where our work should be applied and how it can be continued.

# Chapter 6

# Conclusions

In this final chapter we will be discussing the findings we did with our experiments and propose ideas for future work that may directly use this work as a starting point.

## 6.1 Main Findings

In this work we did experiments in three connected but different objects of work in supervised class imbalanced datasets, features and SVM. The results we obtained on the unbalanced data proved that normal models do not achieve good results for the minority class and that by using models that allow the use of weights this problem can be solved or at least mitigated.
Part of our work is focused solely in achieving better results by the correct use of the features, we concluded that for ROI classification the type of feature we use can greatly change the final result of a model, also we proved that all features can be ranked by level of interest and that by removing those of lower rank allows to produce better results. The main objective of this work is to show that MKL can get better results than SVM. The results we obtained for the DDSM dataset proved exactly that if the same sets of folds and the same methodology is used, SVM alone is much worse than MKL in all the metrics we used for comparison.

## 6.2 Future Work

We would like to use our classifier to create an application that with the supervision of a domain specialist or autonomously can detect areas of interest, extract features from them and do the classification as malignant or benign of tumors in mammograms. To achieve this there are many new problems in several areas that must be solved, because a system like that will require graphical interface, algorithms for computer vision and algorithms for classification and others. There is also space to explore new combinations of kernels and features and how can they be optimally used, this can be an important topic since the number of possible combinations of kernels and features is always quadratic, for example if 5 different kernels are used and the data

contains only 5 features we have more than $5^2$ different ways to combine each feature with a kernel. For higher numbers this could be unfeasible.
A very discussed topic in this area is what family of features are used, because of that, we think that a work focused in experimenting and ranking all the groups of features can be also another topic for exploration since there are works that show up to 264 different single features.
From all the works we cite and others, we concluded that this area of research is in need of a standard for comparing all the developed works. We suggest the creation of a single repository with a selection of images from several datasets and another repository with objects describing ROI with all the features already extracted, if all the works use that same dataset and list what objects are using it will be easier to compare the results between all of them. Also there is space for a work focused in exploring the possibility of ranking all the groups of features according to the different methods, mainly because features that work better for SVM may not have the same performance in a Bayes Network, and there is no work that concludes what group of features presents better results across all methods. We also propose that all the feature selectors that have been shown in several works, should be tested under the same dataset and results compared.

## 6.3 Conclusion

We presented two models, based on Multiple Kernel Learning, for classifying malignant findings in mammography images. The main contribution of this work is to handle the inherently heterogeneous data that usually come from the medical domain. We apply MKL with an SVM classifier in order to discriminate between malignant and benign findings. Our MKL learning uses a cost-sensitive model in order to focus on the malignant cases (smaller class), and reduce the error in this class. Although we focus on the malignant class, our model performs quite well on the class of benign cases, when compared with other works in the literature and when compared with the clinical performance. In the test set, our weighted model reaches a recall for the malignant class of up to 100% while giving 50% recall for the benign class, which means that this model can be applied to real life applications missing less True Positives than clinical practice. We believe that further exploration of other kernels and features combinations could produce even better results.

# Bibliography

American college of radiology. URL https://www.acr.org/~/media/ACR/Documents/PDF/QualitySafety/Resources/BIRADS/Posters/BIRADS-Reference-Card_web_F.pdf?la=en. Accessed: 2017-01-4. 6

National cancer institute. URL https://www.cancer.gov/types/breast/mammograms-fact-sheet. 1

R. Aarthi, K. Divya, N. Komala, and S. Kavitha. Application of feature extraction and clustering in mammogram classification using support vector machine. In *2011 Third International Conference on Advanced Computing*, pages 62–67, Dec 2011. doi: 10.1109/ICoAC.2011.6165150. 25, 30

Samuel G Armato, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Medical physics*, 38(2): 915–931, 2011. 29

S Audithan et al. Analysis of different types of entropy measures for breast cancer diagnosis using ensemble classification. *Biomedical Research*, 28(7), 2017. 28, 31

Gustavo Barbosa Augusto. Computer Aided Diagnosis for Breast Cancer Detection. Master's thesis, Department of Computer Science, Faculty of Sciences, University of Porto, Porto, Portugal, December 2014. 1, 2, 21, 25, 26, 31, 32, 33, 53

William E. Barlow, Chen Chi, Patricia A. Carney, Stephen H. Taplin, Carl D'Orsi, Gary Cutter, R. Edward Hendrick, and Joann G. Elmore. Accuracy of screening mammography interpretation by characteristics of radiologists. *Journal of the National Cancer Institute*, 96(24):1840–1850, 12 2004. ISSN 0027-8874. doi: 10.1093/jnci/djh333. 26

BCDR. Breast cancer digital repository. http://bcdr.inegi.up.pt/, 2017. 35

Asa Ben-Hur and Jason Weston. A user's guide to support vector machines. *Data mining techniques for the life sciences*, pages 223–239, 2010. 19

Michael R Berthold, Christian Borgelt, Frank Höppner, and Frank Klawonn. *Guide to intelligent data analysis: how to intelligently make sense of real data*. Springer Science & Business Media, 2010. 10, 14, 18

Shradhananda Beura. *Development of Features and Feature Reduction Techniques for Mammogram Classification*. PhD thesis, 2016. 29, 31

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387310738. 19

Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984. 16

Peng Cao, Xiaoli Liu, Jian Zhang, Wei Li, Dazhe Zhao, Min Huang, and Osmar Zaiane. A l2, 1 norm regularized multi-kernel learning for false positive reduction in lung nodule CAD. *Computer Methods and Programs in Biomedicine*, 140:211 – 231, 2017. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2016.12.007. URL http://www.sciencedirect.com/science/article/pii/S0169260716304369. 29, 31

Caret. Classification and regression training. https://cran.r-project.org/web/packages/caret/index.html. 44

S. Chakraborty, M. K. Bhowmik, A. K. Ghosh, and T. Pal. Automated edge detection of breast masses on mammograms. In *2016 IEEE Region 10 Conference (TENCON)*, pages 1241–1245, Nov 2016. doi: 10.1109/TENCON.2016.7848209. 8

Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, May 2011. ISSN 2157-6904. doi: 10.1145/1961189.1961199. URL http://doi.acm.org/10.1145/1961189.1961199. 19, 42

O. Chapelle, P. Haffner, and V. N. Vapnik. Support vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks*, 10(5):1055–1064, Sep 1999. ISSN 1045-9227. doi: 10.1109/72.788646. 21

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Learning non-linear combinations of kernels. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 396–404. Curran Associates, Inc., 2009. URL http://papers.nips.cc/paper/3692-learning-non-linear-combinations-of-kernels.pdf. 22

Robert G Cowell, Philip Dawid, Steffen L Lauritzen, and David J Spiegelhalter. *Probabilistic networks and expert systems: Exact computational methods for Bayesian networks*. Springer Science & Business Media, 2006. 18

Anneleen Daemen, Dirk Timmerman, Thierry Van den Bosch, Cecilia Bottomley, Emma Kirk, Caroline Van Holsbeke, Lil Valentin, Tom Bourne, and Bart De Moor. Improved modeling of clinical data with kernel methods. *Artif Intell Med*, 54(2):103–114, Feb 30th 2012. Epub 2011 Nov 30. 25

DDSM. Terminology, 2017. URL http://marathon.csee.usf.edu/Mammography/DDSM/ddsm_terminology.html. 38

T. Deserno, M. Soiron, J. Oliveira, and A. Araujo. Towards computer-aided diagnostics of screening mammography using content-based image retrieval. In *2011 24th SIBGRAPI Conference on Graphics, Patterns and Images*, pages 211–219, Aug 2011. doi: 10.1109/SIBGRAPI.2011.40. 28

S Devisuganya and RC Suganthe. Optimized feature selection for breast cancer detection. *IIOAB JOURNAL*, 7(9):825–835, 2016. 29, 31

A. P. Dhawan, Y. Chitre, and C. Kaiser-Bonasso. Analysis of mammographic microcalcifications using gray-level image structure features. *IEEE Transactions on Medical Imaging*, 15(3): 246–259, Jun 1996. ISSN 0278-0062. doi: 10.1109/42.500063. 25

Carl J D'Orsi. *ACR BI-RADS Atlas: Breast Imaging Reporting and Data System*. American College of Radiology, 2013. 6

Issam El-Naqa, Yongyi Yang, Miles N Wernick, Nikolas P Galatsanos, and Robert M Nishikawa. A support vector machine approach for detection of microcalcifications. *IEEE TRANSACTIONS ON MEDICAL IMAGING*, 21(12), 2002. 2

Issam El-Naqa, Yongyi Yang, Nikolas P Galatsanos, Robert M Nishikawa, and Miles N Wernick. A similarity learning approach to content-based image retrieval: Application to digital mammography. *IEEE TRANSACTIONS ON MEDICAL IMAGING*, 23(10):1233, 2004. 2

Fabián Rodrigo Narvárez Espinoza. Caracterizacion de patrones anormales en mamografias. Doctor en Ingeniería - Ingeniería Mecánica y Mecatrónica. Caracterizacion de Patrones Anormales en Mamografias Area de Investigación: Diagnostico asistido por computador, Procesamiento digital de Imágenes médicas, Octubre 2016. URL http://www.bdigital.unal.edu.co/56605/. 28, 31

Pedro Ferreira, Nuno A Fonseca, Inês Dutra, Ryan Woods, and Elizabeth Burnside. Predicting malignancy from mammography findings and image–guided core biopsies. *International journal of data mining and bioinformatics*, 11(3):257–276, 2015. 25, 26, 31

Carlos E. Galván-Tejada, Laura A. Zanella-Calzada, Jorge I. Galván-Tejada, José M. Celaya-Padilla, Hamurabi Gamboa-Rosales, Idalia Garza-Veloz, and Margarita L. Martinez-Fierro. Multivariate feature selection of image descriptors data for breast cancer with computer-assisted diagnosis. *Diagnostics*, 7(1), 2017. 29

Dhiren Ghosh and Andrew Vogt. Outliers: An evaluation of methodologies. In *Joint Statistical Meetings*, pages 3455–3460. American Statistical Association San Diego, CA, 2012. 42

Mehmet Gonen and Ethem Alpayd. Multiple kernel learning algorithms. *J. Mach. Learn. Res.*, 12:2211–2268, July 2011. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1953048.2021071. 19, 22, 26, 42

Qiong Gu, Li Zhu, and Zhihua Cai. Evaluation measures of the classification performance of imbalanced data sets. *Computational intelligence and intelligent systems*, pages 461–471, 2009. 13

Q. Guo, Y. Qu, A. Deng, and L. Yang. A new fuzzy-rough feature selection algorithm for mammographic risk analysis. In *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, pages 934–939, Aug 2016a. doi: 10.1109/FSKD.2016.7603303. 30

Y. Guo, X. Wang, Z. Yang, D. Wang, and Y. Ma. Improved saliency detection for abnormalities in mammograms. In *2016 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 786–791, Dec 2016b. doi: 10.1109/CSCI.2016.0153. 27, 31

Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009. 51

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. 26, 36

Michael Heath, Kevin Bowyer, Daniel Kopans, Richard Moore, and W Philip Kegelmeyer. The digital database for screening mammography. In *Proceedings of the 5th international workshop on digital mammography*, pages 212–218. Medical Physics Publishing, 2000. 26, 36

C. Hiba, Z. Hamid, and A. Omar. An improved breast tissue density classification framework using bag of features model. pages 405–409, Oct 2016. doi: 10.1109/CIST.2016.7805081. 28

Solveig Hofvind, Antonio Ponti, Julietta Patnick, Nieves Ascunce, Sisse Njor, Mireille Broeders, Livia Giordano, Alfonso Frigerio, and Sven Törnberg. False-positive results in mammographic screening for breast cancer in europe: a literature review and survey of service screening programmes. *Journal of Medical Screening*, 19(suppl 1):57–66, 2012. doi: 10.1258/jms.2012.012083. URL http://msc.sagepub.com/content/19/suppl_1/57.abstract. 1

Shujie Hou, Robert C. Qiu, Zhe Chen, and Zhen Hu. SVM and dimensionality reduction in cognitive radio with experimental validation. *CoRR*, abs/1106.2325, 2011. URL http://arxiv.org/abs/1106.2325. 21

Idil Isikli Esener, Semih Ergin, and Tolga Yuksel. A new feature ensemble with a multistage classification scheme for breast cancer diagnosis. *Journal of Healthcare Engineering*, 2017, 2017. 28

John T. James. A new, evidence-based estimate of patient harms associated with hospital care. *Journal of Patient Safety*, 9:122–128, 9 2013. doi: 10.1097/PTS.0b013e3182948a69. 1

A. A. Khan, M. Khan, and A. S. Arora. Automatic detection of malignant neoplasm from mammograms. In *2015 Science and Information Conference (SAI)*, pages 292–297, July 2015. doi: 10.1109/SAI.2015.7237158. 8

I. Kitanovski, B. Jankulovski, I. Dimitrovski, and S. Loskovska. Comparison of feature extraction algorithms for mammography images. In *Image and Signal Processing (CISP), 2011 4th International Congress on*, volume 2, pages 888–892, Oct 2011. doi: 10.1109/CISP.2011.6100285. 2

Daniel B Kopans. *Breast imaging*. Lippincott Williams & Wilkins, 2007. 6

S. Kowsalya and D. S. Priyaa. An integrated approach for detection of masses and macro calcification in mammogram images using dexterous variant median fuzzy c-means algorithm. In *2016 10th International Conference on Intelligent Systems and Control (ISCO)*, pages 1–6, Jan 2016a. doi: 10.1109/ISCO.2016.7727146. 8

S. Kowsalya and D. Shanmuga Priyaa. An adaptive behavioral learning technique based bilateral asymmetry detection in mammogram images. *Indian Journal of Science and Technology*, 9(S1), 2016b. ISSN 0974 -5645. URL http://52.172.159.94/index.php/indjst/article/view/103646. 30

S. Mohan Kumar and G. Balakrishnan. Wavelet and symmetric stochastic neighbor embedding based computer aided analysis for breast cancer. *Indian Journal of Science and Technology*, 9 (47), 2016. ISSN 0974 -5645. URL http://52.172.159.94/index.php/indjst/article/view/106512. 29, 31

Finn Kuusisto, Inês Dutra, Mai Elezaby, Eneida A. Mendonça, Jude Shavlik, and Elizabeth Burnside. Leveraging expert knowledge to improve machine-learned decision support systems. In *Summit on Clinical Research Informatics within AMIA 2015 Joint Summits on Translational Science*, AMIA Jt Summits Transl Sci Proc, page 87–91, San Francisco, CA, USA, March 2015. AMIA (American Medical Informatics Association), AMIA (American Medical Informatics Association). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4525246/. 32

X. Liu, J. Liu, D. Zhou, and J. Tang. A benign and malignant mass classification algorithm based on an improved level set segmentation and texture feature analysis. pages 1–4, June 2010. ISSN 2151-7614. doi: 10.1109/ICBBE.2010.5518284. 27, 53

X. Liu, L. Zhai, and T. Zhu. Recognition of architectural distortion in mammographic images with transfer learning. In *2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 494–498, Oct 2016a. doi: 10.1109/CISP-BMEI.2016.7852761. 27, 31

X. Liu, L. Zhai, T. Zhu, and Z. Yang. Architectural distortion recognition based on a subclass technique and the sparse representation classifier. In *2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 422–426, Oct 2016b. doi: 10.1109/CISP-BMEI.2016.7852748. 29, 31

Xiaoming Liu, Bo Li, Jun Liu, Xin Xu, and Zhilin Feng. *Mass Diagnosis in Mammography with Mutual Information Based Feature Selection and Support Vector Machine*, pages 1–8. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-31576-3. doi: 10.1007/978-3-642-31576-3_1. URL http://dx.doi.org/10.1007/978-3-642-31576-3_1. 2

F. Ma, L. Yu, M. Bajger, and M. J. Bottema. Mammogram mass classification with temporal features and multiple kernel learning. In *Digital Image Computing: Techniques and Applications (DICTA), 2015 International Conference on*, pages 1–7, Nov 2015. doi: 10.1109/DICTA.2015.7371282. 2

Hamed Masnadi-Shirazi, Nuno Vasconcelos, and Arya Iranmehr. Cost-sensitive support vector machines. *CoRR*, abs/1212.0975, 2012. URL http://arxiv.org/abs/1212.0975. 20, 42

A. Melouah and H. F. Merouani. Hierarchical segmentation of digital mammography by agents competition. In *Digital Information Management, 2008. ICDIM 2008. Third International Conference on*, pages 442–447, Nov 2008. doi: 10.1109/ICDIM.2008.4746776. 2

MIAS. The mini-mias database of mammograms. http://peipa.essex.ac.uk/info/mias.html, 2017. 26

Marina Milosevic, Zeljko Jovanovic, and Dragan Jankovic. A comparison of methods for three-class mammograms classification. *Technology and Health Care*, (Preprint):1–14, 2017. 28, 31

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. ISSN 1076-6332. doi: https://doi.org/10.1016/j.acra.2011.09.014. URL http://www.sciencedirect.com/science/article/pii/S107663321100451X. 26

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. ISSN 1861-6429. doi: 10.1007/s11548-013-0838-2. URL http://dx.doi.org/10.1007/s11548-013-0838-2. 1, 8, 26, 31, 35

Daniel Cardoso Moura, Miguel Angel Guevara López, Pedro Cunha, Naimy González de Posada, Raúl Ramos Pollan, Isabel Ramos, Joana Pinheiro Loureiro, Inês C. Moreira, Bruno M. Ferreira de Araújo, and Teresa Cardoso Fernandes. Benchmarking datasets for breast cancer computer-aided diagnosis (cadx). In José Ruiz-Shulcloper and Gabriella Sanniti di Baja, editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 18th Iberoamerican Congress, CIARP 2013, Havana, Cuba, November 20-23, 2013, Proceedings, Part I*, pages 326–333. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-41822-8. doi: 10.1007/978-3-642-41822-8_41. URL http://dx.doi.org/10.1007/978-3-642-41822-8_41. 26, 31, 35

Marimuthu Muthuvel, Balakumaran Thangaraju, and Gowrishankar Chinnasamy. Microcalcification cluster detection using multiscale products based hessian matrix via the tsallis thresholding scheme. *Pattern Recognition Letters*, pages –, 2017. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2017.05.002. URL http://www.sciencedirect.com/science/article/pii/S0167865517301447. 27

Nadim Nachar et al. The mann-whitney u: A test for assessing whether two independent samples come from the same distribution. *Tutorials in Quantitative Methods for Psychology*, 4(1):13–20, 2008. 11

Fabián Narváez, Gloria Díaz, Cesar Poveda, and Eduardo Romero. An automatic BI-RADS description of mammographic masses by fusing multiresolution features. *Expert Systems with Applications*, 74:82 – 95, 2017. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2016.11.031. URL http://www.sciencedirect.com/science/article/pii/S0957417416306662. 29, 31

R Nithya and B Santhi. Computer-aided diagnosis system for mammogram density measure and classification. *Biomedical Research*, 28(6), 2017. 27, 30

A. Oliver, J. Freixenet, R. Marti, J. Pont, E. Pérez, E. R. E. Denton, and R. Zwiggelaar. A novel breast tissue density classification methodology. *IEEE Transactions on Information Technology in Biomedicine*, 12(1):55–65, Jan 2008. ISSN 1089-7771. doi: 10.1109/TITB.2007.903514. 2

Noel Pérez Pérez, Miguel A Guevara López, Augusto Silva, and Isabel Ramos. Improving the mann–whitney statistical test for feature selection: An approach in breast cancer diagnosis on mammography. *Artificial intelligence in medicine*, 63(1):19–31, 2015. 11

Foster J Provost, Tom Fawcett, et al. Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions. In *KDD*, volume 97, pages 43–48, 1997. 13

A. Qayyum and A. Basit. Automatic breast segmentation and cancer detection via SVM in mammograms. In *2016 International Conference on Emerging Technologies (ICET)*, pages 1–6, Oct 2016. doi: 10.1109/ICET.2016.7813261. 8

Yuchen Qiu, Shiju Yan, Rohith Reddy Gundreddy, Yunzhi Wang, Samuel Cheng, Hong Liu, and Bin Zheng. A new approach to develop computer-aided diagnosis scheme of breast mass classification using deep learning technology. *Journal of X-Ray Science and Technology*, (Preprint):1–13, 2017. 28, 32

P. Rahmati and A. Ayatollahi. Maximum likelihood active contours specialized for mammography segmentation. In *2009 2nd International Conference on Biomedical Engineering and Informatics*, pages 1–4, Oct 2009. doi: 10.1109/BMEI.2009.5305011. 2

Raúl Ramos-Pollán, Miguel Angel Guevara-López, Cesar Suárez-Ortega, Guillermo Díaz-Herrero, Jose Miguel Franco-Valiente, Manuel Rubio-del Solar, Naimy González-de Posada, Mario Augusto Pires Vaz, Joana Loureiro, and Isabel Ramos. Discovering mammography-based machine learning classifiers for breast cancer diagnosis. *Journal of Medical Systems*, 36(4): 2259–2269, 2012. ISSN 1573-689X. doi: 10.1007/s10916-011-9693-2. URL http://dx.doi.org/10.1007/s10916-011-9693-2. 26, 35

Craig Saunders, Alexander Gammerman, and Volodya Vovk. Ridge regression learning algorithm in dual variables. In *Proceedings of the Fifteenth International Conference on Machine Learning*,

ICML '98, pages 515–521, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 1-55860-556-8. URL http://dl.acm.org/citation.cfm?id=645527.657464. 22

Zobia Suhail, Erika R. E. Denton, and Reyer Zwiggelaar. Tree-based modelling for the classification of mammographic benign and malignant micro-calcification clusters. *Multimedia Tools and Applications*, pages 1–14, 2017. ISSN 1573-7721. doi: 10.1007/s11042-017-4522-3. URL http://dx.doi.org/10.1007/s11042-017-4522-3. 28, 31, 53

Deepti Tamrakar and Kapil Ahuja. Density-wise two stage mammogram classification using texture exploiting descriptors. *CoRR*, abs/1701.04010, 2017. URL http://arxiv.org/abs/1701.04010. 30, 31

Maxine Tan, Faranak Aghaei, Yunzhi Wang, and Bin Zheng. Developing a new case based computer-aided detection scheme and an adaptive cueing method to improve performance in detecting mammographic lesions. *Physics in Medicine and Biology*, 62(2):358, 2016. 30, 31

Soumaya Trabelsi Ben Ameur, Florence Cloppet, Sellami Dorra, and Laurent Wendling. Choquet Integral based Feature Selection for Early Breast Cancer Diagnosis from MRIs. In *ICPRAM*, volume Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods, pages 351 – 358, Rome, Italy, February 2016. doi: 10.5220/0005754703510358. URL https://hal.archives-ouvertes.fr/hal-01391970. 31

S Venkatalakshmi and J Janet. Classification of mammogram abnormalities using pseudo zernike moments and SVM. *International Journal of Image, Graphics and Signal Processing*, 9(4):30, 2017. 28, 31

Nisar Wani and Khalid Raza. Multiple kernel learning approach for medical image analysis. *bioRxiv*, 2017. doi: 10.1101/121509. URL http://biorxiv.org/content/early/2017/03/29/121509. 29, 31

Liyang Wei, Yongyi Yang, Robert M Nishikawa, and Yulei Jiang. A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications. *IEEE transactions on medical imaging*, 24(3):371–380, 2005. 2, 25, 30

Liyang Wei, Yongyi Yang, and Robert M Nishikawa. Microcalcification classification assisted by content-based image retrieval for breast cancer diagnosis. *Pattern recognition*, 42(6):1126–1132, 2009. 2

Liu zianzhuang LI Wenqin. The automatic thresholding of gray-level pictures via two-dimensional otsu method [j]. *Acta Automatica Sinica*, 1:015, 1993. 8

Peter H Westfall and S Stanley Young. *Resampling-based multiple testing: Examples and methods for p-value adjustment*, volume 279. John Wiley & Sons, 1993. 14

X. Yang, H. Peng, and M. Shi. SVM with multiple kernels based on manifold learning for breast cancer diagnosis. In *Information and Automation (ICIA), 2013 IEEE International Conference on*, pages 396–399, Aug 2013. doi: 10.1109/ICInfA.2013.6720330. 2

P. C. Yuen, Y. Y. Wong, and C. S. Tong. Contour detection using enhanced snakes algorithm. *Electronics Letters*, 32(3):202–204, Feb 1996. ISSN 0013-5194. doi: 10.1049/el:19960163. 8

T. Zare, M. T. Sadeghi, and H. R. Abutalebi. A comparative study of multiple kernel learning approaches for SVM classification. In *Telecommunications (IST), 2014 7th International Symposium on*, pages 84–89, Sept 2014. doi: 10.1109/ISTEL.2014.7000674. 2

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*, RACS '11, pages 232–237, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-1087-1. doi: 10.1145/2103380.2103426. URL http://doi.acm.org/10.1145/2103380.2103426. 1