INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE MONTERREY
CAMPUS MONTERREY

SCHOOL OF ENGINEERING AND INFORMATION TECHNOLOGIES
GRADUATE PROGRAMS

DOCTOR OF PHILOSOPHY
in
INFORMATION TECHNOLOGIES AND COMMUNICATIONS
MAJOR IN INTELLIGENT SYSTEMS

A Case-Based Reasoning Methodology Applied to the Classification of Microcalcification Clusters and Masses in Breast Cancer Computer-Aided Detection

By

Edén Alejandro Alanís Reyes

MAY 2014

A Case-Based Reasoning Methodology Applied to the Classification of Microcalcification Clusters and Masses in Breast Cancer Computer-Aided Detection

A dissertation presented by

Edén Alejandro Alanís Reyes

Submitted to the
Graduate Programs in Mechatronics and Information Technologies
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Information Technologies and Communications
Major in Intelligent Systems

Thesis Committee:

Dr. Hugo Terashima Marín	-	Tecnológico de Monterrey
Dr. Santiago Enrique Conant Pablos	-	Tecnológico de Monterrey
Dr. Sigfrido Iglesias Gonzáles	-	Tecnológico de Monterrey
Dr. José Tamez Peña	-	Tecnológico de Monterrey
Dr. Oscar Cordón	-	European Centre for Soft Computing
Dr. Sergio Damas	-	European Centre for Soft Computing

Instituto Tecnológico y de Estudios Superiores de Monterrey
Campus Monterrey
May 2014

Instituto Tecnológico y de Estudios Superiores de Monterrey
Campus Monterrey

Division of Mechatronics and Information Technologies
Graduate Program

The committee members, hereby, certify that have read the dissertation presented by Edén Alejandro Alanís Reyes and that it is fully adequate in scope and quality as a partial requirement for the degree of **Doctor of Philosophy in Information Technologies and Communications**, with a major in **Intelligent Systems**.

Dr. Hugo Terashima Marín
Tecnológico de Monterrey
Principal Advisor

Dr. Santiago Enrique Conant Pablos
Tecnológico de Monterrey
Committee Member

Dr. Sigfrido Iglesias Gonzáles
Tecnológico de Monterrey
Committee Member

Dr. José Tamez Peña
Tecnológico de Monterrey
Committee Member

Dr. Oscar Cordón
European Centre for Soft Computing
Committee Member

Dr. Sergio Damas
European Centre for Soft Computing
Committee Member

Dr. César Vargas Rosales
Director of the Doctoral Programme in Information Technologies and Communications
School of Engineering and Information Technologies

Copyright Declaration

I, hereby, declare that I wrote this dissertation entirely by myself and, that, it exclusively describes my own research.

Edén Alejandro Alanís Reyes
Monterrey, N.L., México
May 2014

Dedication

Acknowledgements

A Case-Based Reasoning Methodology Applied to the Classification of Microcalcification Clusters and Masses in Breast Cancer Computer-Aided Detection

by

Edén Alejandro Alanís Reyes

Abstract

This PhD Dissertation provides a thorough description of the Doctoral research project that the author conducted within the School of Engineering and Information Technologies, under the supervision of Professor Hugo Terashima, starting on august 2010.

The main objective of this research project is to design a Case-Based Reasoning (CBR) classification methodology that can be applied to the analysis and diagnosis of breast cancer lesions (masses and microcalcification clusters, specifically), that are also automatically detected in digital mammograms by our framework, as a tool to enhance the diagnosis accuracy in the earliest stages of that disease, when it can still be cured. This framework is intended to to enhance the performance of a given classifier, as compared to the scenario in which the algorithm is not used within our methodology.

We designed a k-nearest neighbour similarity-search combined with a binary genetic algorithm, as a feature selection mechanism, to retrieve similar cases from a historical database that serves as the ground truth of the system, provided that it contains a set of cases that were already revised and validated by an expert. Those similar cases are then used to train a classifier to discriminate between benign and malignant lesions encountered in the query case.

In our experiments we present an empirical study in which we analyze the fitness of six different similarity metrics applied in a k-nearest neighbors similarity-search on our datasets, for case retrieval. Afterwards, we conducted pairwise statistical tests to determine that linear correlation is the best similarity metric for our datasets. Additionally, we explore the performance of our proposed CBR approach in assessing the malignancy of detected lesions by using four different classifiers: k-NN, neural network, SVM and Linear Discriminant Analysis, and also compared its performance against classification results that were obtained with the same classifiers under a traditional CAD pipeline. We conclude that the implementation of our methodology results in an significant increase of the classification performance of the considered algorithms. Finally, we present a study that compares our results against the performance of related classification approaches, including classifier ensembles and techniques for dealing with the class-imbalance problem.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Cancer is a general term used to describe more than 100 different types of uncontrolled growth of abnormal cells; cancer cells divide and reproduce abnormally, invading and destroying surrounding tissue and typically leave the site where they first originated and move to other parts of the body, beginning new cancerous tumors.

According to the Encyclopedia of Breast Cancer [94], it is one of the most common types of cancer among women in the United States, where it is diagnosed in more than 180,000 women and the typical sign of breast cancer is a lump in the breast or under the arm. In general terms, this disease is found in 25 of every 100,000 persons, of which 99% are women. The exact causes of breast cancer are not known, but studies show that older women are more likely to develop that disease, since more cancers occur in women over age 50 and specially in women over age 60. Furthermore, there are several other factors that research has shown to increase the risks of having breast cancer, such as race, family history, alcohol consumption or obesity, among others [94].

Breast cancer rates are the highest in industrialized countries, where about 1.2 million new diagnoses and over 500,000 deaths are reported worldwide in a yearly basis. Since 1980, breast cancer rates have increased by 26 percent. In the United States, one in eight women will develop breast cancer in her lifetime, while in the United Kingdom, it is one in 10 to 12, and in Australia one in 14. In Hong Kong, one in 24 women has this disease each year. In Japan, the incidence is lower than in the United States, but is was reported that it doubled between 1960 and 1980.

However, while breast cancer incidence is increasing, death rates in industrialized countries are declining. In fact, countries with the highest number of new cases annually have seen the greatest decline in mortality rates (United States, Canada, Austria, Germany, the United Kingdom, Australia and Sweden). On the other hand, countries that have lower incidence rates are experiencing an increasing mortality rate; often, these are developing countries.

In the United States alone, an estimated of 215,990 women were diagnosed with invasive breast cancer and 59,390 were diagnosed with *ductal carcinoma in situ* (a noninvasive cancer contained in the milk ducts), and 40,110 died of breast cancer, in 2004. Moreover, about 3 million american women are living with breast cancer; two million of them have been diagnosed with the disease and one million have the disease but do not yet know it.

This health problem is much more serious in developing countries, where an appropriate public health infrastructure is typically not present or is not covered by social security, leaving the population with few options to get medical treatment and resulting in a very small percentage of patients being able to undergo clinical breast cancer examinations. According to Knaul et al [54], in Mexico, for instance, breast cancer accounts for more deaths than cervical cancer since 2006 and it is the second cause of death for women aged 30 to 54 from all socio-economic groups. In spite of the fact that Mexico's social

health insurance covers breast cancer treatment, its infrastructure reaches approximately 40-45% of the entire population. There is a lack of diagnosis services and medical interventions for early detection are limited, particularly mammography. In 2006, only 22% of women between 40 to 69 years old reported having undergone a mammogram in the past year; also, the vast majority of cases are self-detected and only 10% of them belong to the earliest stage of development, as compared to 50% in the United States.

Since there is no effective prevention, early diagnosis and effective treatment are the only options to decrease mortality rates due to this disease. As a prevention mechanism, women should have regular breast exams, that can include physical exams, mammograms, ultrasonography, or many others. Currently, screening mammograms are the best tool available for finding breast cancer early because with them physicians can detect lumps or other lesions that are possibly not yet palpable but may be an early evidence of cancer or a precancerous change in the tissue. Mammograms are recommended for young women who have symptoms of breast cancer or have a high risk of it, given their family history, and for women older than 40, as well, even if they have no signs of the disease.

For instance, a mammography may reveal small white spots on the film, which are tiny mineral deposits within the breast tissue called calcifications, or may highlight a suspicious mass, architectural distortions, and asymmetric densities. Computer-aided diagnosis tools use this abnormal findings in order to make a decision whether or not they are malicious. For instance, isolated calcifications have low chances of being malignant, but small calcifications forming a group or cluster of three or more of them have high chances of begin malignant.

The information that a physician may obtain from mammography is of high value to determine the malignancy of a given lesion. However, this screening exam has some limitations in terms of sensitivity and specificity that affect the outcome, which is why it is often followed by other clinical studies that include biopsy, an examination with a different technology (such as a magnetic resonance image), or even a second diagnosis of the same mammograms performed by a different physician with the objective of confirming the initial findings or rejecting/correcting them for further revisions, which is quite inefficient in terms of time and still may not result in an accurate diagnosis.

This research work has the objective of design and develop an artificial intelligence method for the detection of early breast cancer, based on the abnormalities found in digital mammograms; an image retrieval retrieval mechanism will be developed to extract historical cases with known pathologies from a database, in order to be considered in the classification task, and to be displayed to the user as a justification of the system's answer. The classification methodology presented in this document considers the diagnosis of microcalcification clusters and masses, which are the two most common breast cancer lesions.

1.1 Motivation

According to the World Health Organization [71], cancer, in its different forms, is a leading cause of death worldwide: it accounted for 7.6 million deaths, in 2008. Breast cancer is the fifth deadliest type of cancer (after lung, stomach, colorectal and liver cancer) since it accounted for 458,000 deaths in the same year, and it is the most frequent type of cancer among women, globally.

There are different technological advances for early breast cancer detection, which include mammograms, ultrasound, Magnetic Resonance Imaging (MRI), among others; however, mammography is still the most used tool nowadays, specially in developing economies, because its low cost and the great possibilities that they provide for radiologists to conduct a breast cancer diagnosis, since several breast features can be observed.

A mammography is a non-invasive technique which allows for breast cancer diagnosis. It is a

special series of X-rays that show images of the soft tissues of the breast, and it is aimed to help find cancer in early stages, when it can still be cured; findings in mammograms include masses, microcalcifications, architectural distortions and asymmetric densities. Several research efforts have focused in detecting microcalcifications, since they are evidence of breast cancer and they are present in the earliest stages of the disease; therefore, its detection leads to an early diagnosis and gives the patient the opportunity to follow the appropriate medical treatment to cure this disease.

Furthermore, the sensitivity of a mammography (i.e. proportion of positives which are correctly identified as such) is around 79%, which is higher than a physical clinical exam, but lower than a biopsy. On the other hand, the specificity of a mammogram (i.e. proportion of negatives which are correctly identified) is around 90% [68], and is often complemented with other clinical studies that include biopsy, which has a very high specificity, or even a second diagnosis performed by a different radiologist.

This double diagnosis mechanism increases sensitivity in only 15% [22], which means that a very small amount of missed positive cases are detected by a second radiologist. It is a redundant process that results in having two professionals analyzing the exact same cases, when they could be working in different sets of cases and, thus, processing a greater amount of exams. This is why several research efforts have been directed towards the development of Computer-Aided Diagnosis tools, often based on artificial intelligence techniques, which can support the radiologists' work, by performing one analysis, hence, providing a second opinion, that the radiologist can take into account.

It has been shown that radiologist's misinterpretation of the lesion can lead to a greater number of false positives; between 65% to 90% of the biopsies of suspected cancers turn out to be benign [21] and, hence, provides evidence for the fact that there exists a need for the development of methods which are capable of increasing the lesion-detection and cancer-diagnosis accuracy, for physicians to make better decisions regarding follow-up and biopsy. This is specially important if we also consider that humans are susceptible to make mistakes, and their analysis is usually *subjective* and *qualitative*, while computers, on the other hand, provide an *objective* and *quantitative* biomedical image analysis, which leads the medical doctor to a more accurate diagnostic decision [81].

Thus, this research work is mainly motivated by the fact that there is a raising need to design and develop computer-aided tools that can take advantage of advanced classification techniques, that can result in the so-called *Computer-Aided Diagnosis* (CAD) tools, that have recently been widely tested and used, and which have the objective of performing the diagnosis tasks that a radiologist does, in order to provide them with a second opinion, therefore, increasing the accuracy and efficiency of the process.

1.2 Problem Statement and Context

Even though mammography is the standard in early breast cancer screening, they are difficult to analyze and interpret, especially when performing a screening mammography in asymptomatic women, because the probability of encountering an abnormality is low. Thus, a very efficient method must be designed and developed, in order to achieve high diagnostic accuracy in the resulting CAD, taking into account objective methods for the analysis of mammographic features in computer-aided diagnosis of breast cancer.

Several research efforts have been directed in developing CAD systems for the detection and classification of microcalcifications and masses, such as the ones described in [102, 42, 30, 75, 103, 99, 77, 97], just to mention some of the methods that were developed in the last decade. Typically, the processes that these systems describe follow some common steps listed in Figure 1.1. According to [14], the process starts with the acquisition of a mammogram (either digital or digitized), as the

input of the system; later, a preprocessing stage is performed, in order to decrease noisy features in the film such as background labels. Then, signals (objects that resemble breast lesions) are extracted form the mammogram, by using the appropriate image processing techniques. Afterwards, features are extracted from the previous signals; there are two types of features: gradient-based, intensity-based and geometric features (traditionally used by radiologists) and high-order features that may not be as intuitive to radiologists (texture features). The process ends with a signal classification step, in which false signals are separated from the suspect lesions.

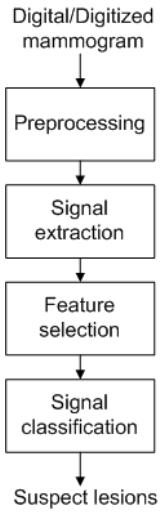


Figure 1.1: Typical CAD stages.

Each stage of most CAD systems uses multiple parameters such as threshold values, filter weights, and region of interest (ROI) sizes. In order to achieve a high performance, these parameters must be tuned optimally; this tuning process is often performed several times, since a modification of any of those parameters may affect the performance of the subsequent stages (they are closely related). A common approach is to empirically select the best parameter values, by trying out different combinations of them, with the aid of automated methods.

Moreover, when several features are extracted, it may be difficult to predict which feature combinations will lead to higher accuracy. Thus, a feature selection step is very often performed, since this is perhaps the most critical stage in a CAD scheme, because including inappropriate features often adversely affects the classification performance.

Another critical stage in this model is the classification of signals, since the researcher has to select an appropriate classification technique, and then be able to design an effective training phase for the selected model, that can result in a highly accurate classification task.

On the other hand, this research work will consider not only the stages of the previously described *traditional*, but also will design and develop additional processes which will be critical to enhance the classification of lesions and, hence, the overall accuracy of the system. One of them is a Content-Based Image Retrieval (CBIR) process, in which a number of already-diagnosed lesions similar to the one currently being analyzed will be gathered from the mammographic database, in order to modify the parameters of the classification algorithm by considering these similar cases and their diagnosis in its training phase. It is important to mention that four classification approaches will be explored in this research work: the k-Nearest Neighbors classifier, the neural network, support vector machines and discriminant analysis, in order to compare them and decide which method fits best as a classifier for this problem.

The overall output of the system will be both the **numerical** answer of the classifier and the **pictorical** output made up with the retrieved images of the lesions. One of the major objectives for doing this is to keep the CAD system from being a *black-box* to the end user, who will typically be an expert radiologist, by providing not only the resulting diagnosis of the classification phase, but also a series of images from similar cases with previously diagnosed pathologies, as a justification of the system's answer. In this way, the radiologist is expected to understand the reasoning of the system.

To do this, a suitable database will be designed and implemented, storing not only mammographic images, but also the *meta-data* necessary to perform queries and retrieve images from it; i.e. to enable the Content-Based Image Retrieval.

Finally, one can see that, in order to design and develop an accurate computer-aided diagnosis system for breast cancer, different challenging issues have to be taken into account. Those issues will be addressed by the present research work, since they represent the problem statement of this Dissertation.

1.3 Objectives

The general objective of this doctoral dissertation is to develop a methodology for the detection and classification of critical breast lesions, such as microcalcification clusters and breast masses, implemented under the Case-Based Reasoning learning paradigm. The classifiers are trained with historical similar cases every time a new lesion has to be classified, providing them with the ability to adapt their parameters based on the features of the new case, thus, being more likely to provide a more accurate diagnosis, as opposite to *static* classifiers that are trained once with representative cases and remain the same from that point on.

To achieve this goal, an in-depth study of several research topics was conducted during the development of this research work, including the Case-Based Reasoning philosophy, machine learning methods, Content-Based Image Retrieval mechanisms that are useful for gathering up relevant data/images from the database, and different ways in which we could combine of all of those techniques into a novel breast-cancer classification methodology.

The particular objectives of this research are:

- To implement a process for the automatic segmentation of potential microcalcification clusters and masses from digital mammographies, based on computer vision and image processing techniques.
- Define a set of features to be extracted from microcalcifications clusters and masses, that best describes their visual content.
- To define an accurate method to select the subset of features that provide the greatest discriminating power to correctly classify suspicious regions of a mammography into benign or malignant lesions.
- To define a set of classifiers that will be used to conduct experiments, considering the inclusion of methods based on different learning approaches.
- Design an appropriate mechanism to retrieve from the database images that contain lesions that are similar to a given query case, using techniques from the Content-Based Image Retrieval literature.
- To determine by experimentation the most appropriate metric to compute the similarity between any two given images, based on their feature vectors.

- To measure the performance of the considered algorithms when they are implemented under a typical/traditional CAD pipeline, in which they are trained with a static subset of a given dataset and a series of *sequential* processes are executed.
- To conduct experimentation to measure the performance that can be acquired by using the considered classifiers under the proposed CBR model.
- To compare the results of both approaches and determine if a better performance is achieved by applying the proposed model.
- To conduct further experimentation to explore the performance of classification algorithms specialized in learning from class-imbalanced data problem, which is an intrinsic issue of medical data.
- To compare performance results achieved with the proposed model against those that are obtained with the considered class-imbalance learning algorithms, in order to provide benchmarking information that can describe how competitive it is.

The utter goal of this research project is to develop a CAD model that can serve as a second opinion for radiologists in the detection and diagnosing of breast cancer. Besides, the model should be designed in a *scalable* way, to allow the integration of modules that can be developed in future research efforts.

1.4 Hypothesis and Research Questions

In this section we present our general basic research hypothesis that drive this research effort. In brief, we believe that:

"It is feasible to develop a classification methodology under a Case-Based Reasoning (CBR) paradigm that can diagnose breast lesions with high accuracy, by taking advantage of the information drawn from similar historical cases with previously diagnosed pathologies, which are used to build classification algorithms that discriminate between benign and malignant cases, leveraging their performance by feeding them only the most relevant information for their classification task".

1.4.1 Research Questions

The path of this research project can be outlined via the following research questions:

- Is there an image processing method that can be used to enhance the resolution of a mammographic image, such that any noise it could contain does not impact negatively the detection and classification of breast abnormalities?
- What image processing methods can we used to develop an image segmentation process that can accurately segment microcalcification clusters from digital mammograms?
- What image processing methods can we used to implement an image segmentation process that can accurately segment masses from digital mammograms?
- Is there a set of features that can provide the classification process with sufficient information for discriminating between real breast cancer lesions and normal tissue?

- What is the most accurate set of features that can provide information about clusters of microcalcifications?
- What is the most appropriate set of features that can be used to describe the visual content of masses?
- What classification algorithms should be considered to be used under our CBR framework, for classifying clusters of calcifications and masses into benign and malignant, ensuring that they can generalize well when facing new cases?
- What is the most adequate way to formulate an *image signature* that can enable a mammogram retrieval process?
- Is there an accurate way to retrieve *similar* cases stored in a mammogram database, based on an input mammogram which acts as a query image?
- How can we measure the similarity between cases?
- What is the best technique we should implement to ensure an accurate retrieval of *similar* cases?
- How can the information drawn from the retrieved similar cases, such as the diagnosed pathology, be used to enhance the performance of a classification algorithm?
- How does the performance of applying our CBR approach compare against a scenario in which it is not applied, i.e. a traditional CAD model?
- How competitive our model is if we compare its performance against other learning paradigms such as classifier ensembles that use techniques for learning from class-imbalanced datasets?

Having all the aforementioned questions as a background, this research project focuses in designing and implementing a CBR-based classification methodology, in which different machine learning algorithms will be tested to evaluate the ability and potential of the CBR framework in making classifiers perform with higher accuracy in discriminating between malignant and benign clusters of microcalcifications and masses. A key component of this CBR approach is the design of an accurate mammogram retrieval mechanism, since the output of this process is going to be used in (1) training the selected classifiers and ensuring that they will perform with a higher accuracy and (2) justifying in a pictorial way the output of the system.

1.5 Solution Overview

Figure 1.2 depicts the proposed CBR model that we have designed to support medical diagnosis of breast cancer lesions. A key component of this model is the database which contains information of historical cases with previously diagnosed pathologies. Considering the feature sets that are extracted from the encountered lesions, our system **retrieves** from the database a subset of k lesions that are similar to the ones found in a query image, by performing a k-NN-based similarity search. We have performed an empirical study in which we evaluated six similarity metrics, in order to determine which one is most suitable for our problem by conducting pairwise statistical tests to compare the performance of a k-NN classifier using all the different metrics.

We performed further classification experiments in which the retrieved cases are **reused** to compute the diagnosis of the encountered lesions, based on four classifiers of interest: the k-nearest neighbors (k-NN) [35], a neural network (NN) [36], support vector machine (SVM) classifier [95] and the

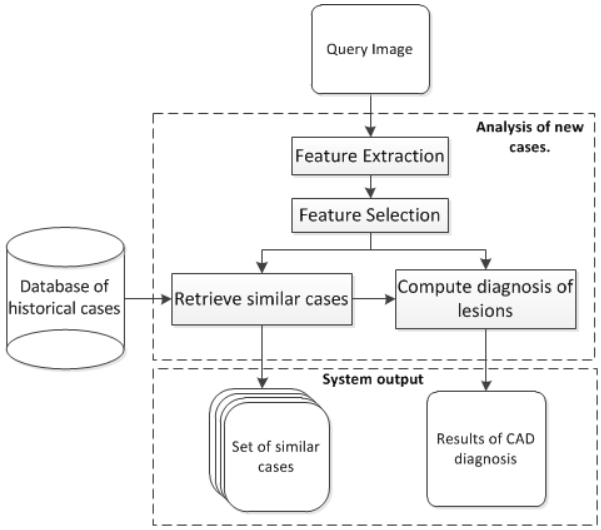


Figure 1.2: Overall model

linear discriminant analysis (LDA) method [35, 34]. We present the performance of these classifiers in terms of area under the receiver operating characteristic (ROC) curve, overall accuracy, sensitivity and specificity.

The system provides two outputs: (1) the query image with overlaid marks in the regions where lesions where found including an assessment of their malignancy, as well as (2) a set of historical cases that have validated pathologies and which are similar to the lesions that were detected in the query case.

In our experiments we study how competitive the proposed model is against both the traditional CAD pipeline that was described in Figure 1.1, which is composed of a series of *sequential* processes, and alternative classification approaches such as classifier ensembles combined with some of the techniques that are used to deal with the class-imbalanced problem, in which one of the classes that should be learned is underrepresented in the dataset, i.e. it contains fewer instances from one class. This issue is often encountered in medical diagnosis problems where there is a large population that undergo this exam but few of them are finally diagnosed as being positive, as it is the case with breast cancer diagnosis.

1.6 Main Contributions

The CBR methodology that is presented in this research work has the following contributions:

- A framework to appropriately build a database of historical cases, which can be used to enable the CBR philosophy. This is critical since the proposed model is based on the implementation of CBR key processes, such as retrieving of historical cases and reusing them; the underlying framework should be designed accordingly.
- A series of image-processing techniques that are able to segment microcalcification clusters and masses. Lesion detection and segmentation is a crucial mechanism since it is from these lesions that the subsequent feature extraction and lesion classification methods are computed.
- A feature extraction stage which is executed upon segmented lesions to characterize their visual

content, in order to enable the implementation of supervised-learning algorithms that will finally assess the malignancy of each lesion.

- An evolutionary wrapper-based feature selection mechanism implementing *diversity injection*, that successfully finds the subset of features that provides a given classifier with the highest discriminant power to discriminate between benign and malignant lesions. Since different classifiers may find different features as being more (or less) relevant for the classification task, we found out that it is very important not only to search the feature space for the most informative ones, but also to tailor this process to each classifier; this resulted in higher performance for all classifiers.

- We propose a statistical way to evaluate and compare the performance of different similarity metrics that are used to retrieve similar cases, using the performance of a k-NN classifier to quantitatively determine which metric fits best for the retrieval task, provided that the difference of precision achieved with it is statistically significant over the others. This was done to ensure high performance in the final classification stage.

- We propose a way of reusing the retrieved cases in the classification of breast cancer lesions. Specifically, we decided to use the retrieved cases as the training set of a classifier algorithm, with the objective of feeding it relevant information only.

- The system shows the retrieved cases to the radiologist, as a way of justifying its output and also with the objective of enabling the physician to take into account the information that can be drawn from the set of historical cases that were used and to combine them with their knowledge as well as with the diagnosis suggested by the system, in order to finally make a more informed and accurate decision.

- The proposed methodology proved to be able to enhance the accuracy of all the considered classifier algorithms using our datasets.

- This classification methodology can be applied to any other datasets, since it follows standard processes used in the machine learning theory and applications.

1.7 Thesis Organization

This document is organized as follows. Chapter 2 presents information related to breast cancer, related risks and symptoms, types of breast cancer, stages of development, techniques that are used for screening and available treatments. Chapter 3 describes what a mammogram is, the process and views of screening breast cancer by mammography, features that are observed in mammograms and features that are related to early breast cancer. Then, Chapter 4 provides information about image mining, including Content-Based Image Retrieval, feature extraction and feature selection techniques, methods used for segmentation of regions of interest and for measuring the similarity between images. Relevant machine learning theory is presented in Chapter 5, focusing on the description of the classification algorithms of interest and the Case-Based Reasoning philosophy. Related methods are described in Chapter , including research works that developed segmentation techniques, feature extraction, feature selection and methods for the classification of breast cancer lesions. Afterwards, Chapter 7 provide a thorough description of the proposed classification methodology, providing with specific details about all the related process within.

Then, Chapter 8 presents classification performance results that describe the accuracy of the considered classifiers without being under the CBR framework, i.e. when they are applied in the

typical CAD pipeline (see Figure 1.1). Chapter 9 introduces the experiments that were conducted to evaluate the performance that can be achieved when we use the considered learning algorithms under the proposed classification methodology. Next, Chapter 10 presents the classification performance that was achieved by classifier ensembles combined with algorithms that are used for learning from class-imbalanced datasets and introduces a discussion about how competitive our model is against these results. Finally, the conclusions of this research effort are presented in Chapter 11, along with future research avenues.

Chapter 2

Breast Cancer

This chapter will provide a description of breast cancer itself, taking into account the factors that contribute to developing the disease and its related symptoms. A description of breast cancer types is also included. Besides, there are different stages of the evolution of breast cancer which are also described and explained, along with the screening techniques commonly used to observe and diagnose it. One of such techniques is the screening mammography, which is in fact the most used one since it presents the radiologist with important information that enables a diagnostic process in early stages, when it can still be cured, thus, a special focus on this topic is provided.

According to the Encyclopedia of Breast Cancer [94], the breast is a mammary gland that produces milk. Each breast has 15 to 20 sections called lobes, which contain many smaller lobules that end in dozens of tiny milk-secreting bulbs embedded in fatty tissue. The lobes, lobules and bulbs are all linked by thin tubes called ducts. These ducts lead to the nipple, located in the center of a dark area of skin called the *areola*. During breast-feeding, milk travels from the lobules into the ducts. There are no muscles in the breast, but muscles lie under each breast and cover the ribs. Each breast also contains blood vessels and vessels carrying colorless fluid called lymph, which lead to small bean-shaped organs called *lymph nodes*. Clusters of lymph nodes are found near the breast in the axilla (under the arm), above the collarbone, and in the chest. Figure 2.1 [67] shows the overall anatomy of the breast.

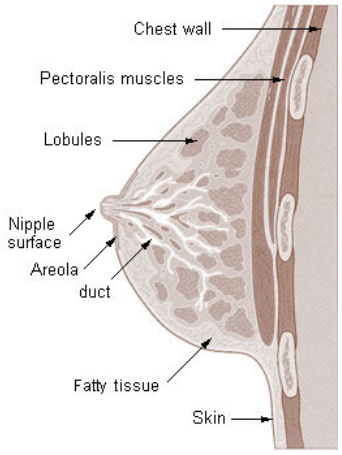


Figure 2.1: Breast Anatomy [67].

The breast is an organ that can have different types of cancers, including lobular, ductal or inflamatory breast cancer.

2.1 Risk factors and Symptoms

The exact causes of breast cancer are not known, but studies show that its risks increase as women get older, since it is uncommon in women under age 35, occurring more often in women over 50 years old and especially high incidence is present in women over age 60. Through several research efforts, scientific consensus have been reached regarding the factors that increase the risks of a woman having breast cancer, which include [94]:

1. ***Race.*** Breast cancer occurs more often in Caucasian women than in African-American or Asian-American women.

2. ***Age.*** This is the most important risk factor. In the United States, chances of a woman having breast cancer is shown in Table 2.1, where we can see that the older a woman is, the greater her risk of having breast cancer.

Age	Probability of having brest cancer
30	1 in 2,525
40	1 in 217
50	1 in 50
60	1 in 24
70	1 in 14
80	1 in 10

Table 2.1: Probability of a woman having breast cancer, according to her age.

3. ***Personal History.*** If a woman had cancer in one breast, there are higher risks for her to have cancer in the opposite one.

4. ***Family History.*** Risks increase if a woman's relatives (mother, sister, daughter) had breast cancer, especially at a young age.

5. ***Breast changes.*** *Lobular carcinoma in situ* (see Section 2.2) may increase a woman's risk for developing cancers.

6. ***Genetic alterations.*** Changes in genes BRCA1 and BRCA2 increase the risk of breast cancer. BRCA1 and BRCA2 are the abbreviations for two *tumor suppressor genes* (BReast CAncer 1 and BReast CAncer 2), that normally help to suppress cell growth.

7. ***Estrogen.*** There is evidence for the fact that the longer a woman is exposed to estrogen, the more likely she is to have breast cancer, whether made by her body, taken as a drug or delivered by a patch. There are important implications of this fact, since it means that, for instance, women who never had children or had the first child after age 30, or took home replacement therapy for long periods, are more likely to develop breast cancer.

8. ***Diethylstilbestrol (DES).*** It is a synthetic form of estrogen that was used between the early 1940s and 1971, to prevent certain complications during pregnancy. Women who took DES are at a slightly higher risk for breast cancer; however, this is apparently not the case for their daughters who were exposed to id before birth.

9. ***Radiation therapy.*** Women whose breasts were exposed to radiation therapy, especially those who were treated with radiation for *Hodgkin's disease*, have greater chances to develop breast cancer.

10. ***Alcohol.*** Evidence suggests that women who have three or more drinks per day have twice the usual chances of developing this disease. However, one to two eight-ounce drinks a day is not associated with an increased risk.

11. ***Obesity.*** Risks increase for postmenopausal obese women. There is no evidence of increased risks for overweight premenopausal women.

Breast cancer, in its early stages, typically does not cause pain or any other symptoms, but the following are present as it grows:

1. A lump or thickening in or near the breast or in the underarm area.
2. A change in the size or shape of the breast.
3. Nipple discharge or tenderness.
4. Inversion of the nipple into the breast.
5. Ridges or pitting of the breast (the skin resembles the skin of an orange).
6. A change in the appearance or texture of the skin of the breast, areola or nipple.

2.2 Types of Breast Cancer

There are several types of lesions or tumors that may develop within different areas of the breast, most of them being the result of non-cancerous changes. Breast cancer types can be broadly classified into *invasive* or *in situ*; the invasive ones are those in which cancer cells break through the duct and lobular wall and invade the surrounding fatty and connective tissues of the breast, while *in situ* refers to the non-invasive breast cancer types in which cancer cells are confined to the ducts and do not invade surrounding fatty and connective tissues (cancer has not spread past the area where it initially developed). According to the Imaginis Corporation [63] and the Encyclopedia of Breast Cancer [94], the different forms of breast cancer that may be found in women are:

1. ***Ductal carcinoma in situ (DCIS).*** This is the most common type of non-invasive breast cancer, accounting for 90% of non-invasive ones. However, it accounts for 25% of all breast cancer diagnoses. It is confined to the ducts of the breast, i.e., there are abnormal cells that are growing in the lining of a duct. With this type of precancer, the abnormal cells have not spread beyond the duct to invade the surrounding breast tissue and so they do not spread to lymph nodes or other organs. DCIS is often first detected on mammogram as microcalcifications (tiny calcium deposits). With early detection, the five-year survival rate for DCIS is nearly 100%, provided that the cancer has not spread past the milk ducts to the fatty breast tissue or any other regions of the body. There are several different types of DCIS. For example, ductal comedocarcinoma refers to DCIS with necrosis (areas of dead or degenerating cancer cells). It is also called *intraductal carcinoma*, and women with untreated DCIS are at an increased risk of invasive breast cancer, as over time DCIS may become invasive. The treatment of DCIS is the same as for invasive cancer: either lumpectomy and radiation or mastectomy (preventive breast removal). Besides, women who have DCIS may want to consider taking tamoxifen to reduce the risk of developing invasive breast cancer.

2. ***Invasive ductal carcinoma (IDC).*** Also know as *infiltrating ductal carcinoma* The most common type of breast cancer, which begins in the cells lining the ducts of the breast. It accounts for 80% to 85% of all breast cancers, which can spread to lymph nodes or other organs; however, if the cancer is less than 1 centimeter in diameter, spread is unlikely. Cells grow only within the duct tube (called intraductal cells), continue along the duct system, which eventually lead to the nipple; between 15 and 20 percent of breast cancers are found at this stage. They can remain in this preinvasive stage for some time and usually these lesions do not cause symptoms, except an occasional bloody discharge from the nipple. Occasionally, these intra-ductal cancer cells may appear as small flecks of calcium on a mammogram.

3. ***Lobular carcinoma in situ (LCIS, lobular neoplasia).*** This is not a cancer; it refers to abnormal cells in the lining of a lobule, and the presence of such cells is a sign that a woman has an increased risk of developing breast cancer in either breast. It is a sharp increase in the number of cells within the milk glands (lobules) of the breast; then, it spreads through the basement membrane into the surrounding breast tissue. This type of cancer can also spread beyond the breast to other parts of the body. About 10 to 15 percent of invasive breast cancers are invasive lobular carcinomas; these tumors feel like thickened areas of the breast instead of lumps. Many physicians do not classify LCIS as breast cancer and often encounter LCIS by chance on breast biopsy while investigating an area of concern. LCIS patients are closely monitored every four months with physician-performed clinical breast exams in addition to receiving yearly mammography. Some women who have LCIS may take tamoxifen, which can reduce the risk of development of breast cancer; others may choose not to have treatment but simply return to the doctor regularly for check-ups. Occasionally, women who have LCIS may decide to have preventive surgery to remove both breasts to try to prevent development of cancer, a technique called prophylactic mastectomy. Since LCIS is a marker of breast cancer risk and not a true cancer, there is no need for surgery or radiation.

4. ***Invasive lobular carcinoma (ILC).*** Also known as *infiltrating lobular carcinoma*, it begins in the milk glands (lobules) of the breast, but often spreads (metastatizes) to other regions of the body. ILC accounts for 10% to 15% of breast cancers.

As mentioned previously, early detection of breast cancer is associated with the highest survival rate, since the patient can still be cured by following the appropriate treatment. This means that diagnosis efforts should be directed to detecting the *in situ* types of breast cancer, when the disease has just began and has not yet spread to other organs, in which case the mortality rate is higher. Having that as a background, and recalling from the definition of the DCIS that it is the most common type of *in situ* breast cancer, it is clear to see why several research efforts have been conducted towards the development of computer-aided DCIS detection systems, providing radiologists with tools to enhance their diagnosis accuracy, which results in a decrease of mortality rates.

Now, there are some other uncommon forms of breast cancer, namely:

1. ***Mucinous carcinoma.*** Also called *colloid carcinoma*, mucinous carcinoma is a rare breast cancer formed by the mucus-producing cancer cells. Women with mucinous carcinoma generally have a better prognosis than women with more common types of invasive carcinoma.

2. ***Phylloides tumor.*** Phylloides tumors can be either benign (non-cancerous) or malignant (cancerous). They develop in the connective tissues of the breast and may be treated by surgical removal. However, they are very rare, since **less than 10 women die** of this type of breast cancer each year in the United States.

3. ***Inflammatory breast cancer.*** This is an uncommon type of locally advanced breast cancer in which the breast looks red and swollen (or inflamed) because cancer cells block the lymph vessels in the skin of the breast. The skin of the breast may also show a pitted appearance called *peau d'orange* (french for "the skin of an orange"). Inflammatory breast cancer **accounts for about 1% of invasive breast cancers**, it is quite likely to spread to lymph nodes or other parts of the body since it generally grows rapidly; it may be diagnosed with swollen lymph nodes but no tumor in the breast. The name for this type of breast cancer was chosen many years ago by doctors who thought the breast tissue was inflamed, but, actually, those skin changes typical of this type of cancer are not due to inflammation but rather to spread of cancer cells within the lymphatic channels of the skin. In order to confirm its diagnosis, a biopsy should be performed, and once diagnosed as positive, the treatment that follows is very aggressive, because surgical removal (even *mastectomy*) does not control it locally, hence, aggressive *chemotherapy* is usually given as a first step to dramatically shrink the cancer and return the skin to a normal appearance. Finally, surgery and/or radiation therapy is then used to remove or destroy the cancer.

4. ***Paget's disease of the nipple.*** This is a rare form of breast cancer that begins in the milk ducts and spreads to the skin of the nipple and areola, in which the breast skin may appear crusted, red, or oozing. A woman's prognosis (expected outcome) may be better if nipple changes are the only sign of the breast disease and no lump is felt. It only **accounts for about 1% of breast cancers**.

5. ***Tubular carcinoma.*** Tubular carcinomas are a special type of invasive breast carcinoma. Women with tubular carcinoma generally have a better prognosis than women with more common types of invasive carcinoma. Tubular carcinomas **account for around 2% of breast cancer diagnoses**.

6. ***Medullary carcinoma.*** It is an invasive breast cancer that forms a distinct boundary between tumor tissue and normal tissue. **Only 5% of breast cancers are medullary carcinoma**.

We can see that these type of breast cancers are very rare and do not account for a considerable amount of cases. Thus, diagnosis efforts are focused mainly in the most typical forms of cancer, previously discussed.

2.3 Stages of Breast Cancer

In most cases, the most important factor is to detect the stage of the breast cancer, since it is useful to select the suitable treatment options. Estimating the stage of this disease is based on the size of the tumor and whether the cancer has spread. In general terms, the smaller the tumor, the better (physicians consider tumors less than two centimeters to be *small*). According to medical consensus, the stages of breast cancer are [94]:

1. ***Stage 0.*** This is sometimes called non-invasive carcinoma or carcinoma in situ.

2. ***Stage I.*** This is an early stage of breast cancer, in which the cancer has spread beyond the lobe or duct and invaded nearby breast tissue. Stage I means that the tumor is no more than about one centimeter across and cancer cells have not spread beyond the breast.

3. ***Stage II.*** This is still considered an early stage of breast cancer. The cancer has spread beyond the lobe or duct and invaded nearby breast tissue. In this stage, there are three possible scenarios:

- The tumor in the breast is less than one centimeter across and the cancer has spread to the lymph nodes under the arm,
- The tumor is between one and two centimeters (with or without spread to the lymph nodes under the arm), or
- The tumor is larger than two centimeters but has not spread to the lymph nodes under the arm.

4. ***Stage III.*** This is also called *locally advanced cancer*. In this stage, the tumor in the breast is large (more than two centimeters across) and the cancer has spread to one of the following parts of the body:
   - To the underarm lymph nodes,
   - To lymph nodes near the breastbone,
   - To other tissues near the breast, or
   - The cancer is extensive in the underarm lymph nodes.
5. ***Stage IV.*** This is metastatic cancer in which the malignancy has spread beyond the breast and underarm lymph nodes to other parts of the body.
6. ***Recurrent.*** This means that the disease has returned in spite of the initial treatment. Most recurrences appear within the first two or three years after treatment, but breast cancer can recur many years later. Cancer that returns only in the area of the surgery is called a *local recurrence*; if the disease returns in another part of the body, the distant recurrence is called *metastatic breast cancer*. The patient may have one type of treatment or a combination of treatments for recurrent cancer.

For a more specific and detailed explanation of breast cancer staging, the reader should refer to the 2010 *Cancer Staging Manual*, prepared by the American Joint Committee on Cancer (AJCC) in [29], which uses the *TNM* system for the classification of malignant tumors. The *TNM* system is a globally accepted method of describing the anatomical extent of cancer.

2.4 Breast Cancer Screening and Diagnosis

Women should have regularly scheduled screening exams, such as mammograms and clinical breast exams. A screening mammogram is the best tool available for finding breast cancer early and can often detect a breast lump before it can be felt. In general terms, diagnosis of breast cancer includes a careful physical exam, personal and family medical history, together with one or more of the following breast exams [94]:

- ***Clinical breast exam.*** The doctor should carefully feel the breast and the tissue around it. Benign lumps often feel different from cancerous ones; the doctor can examine the size and texture of the lump and determine whether the lump moves easily.
- ***Breast Self-Exam.*** The patient herself performs this exam on her own. The objective of this method is to feel possible distortions or swelling on each breast. It is an easy but not very reliable method for finding possible breast cancer; however, experts still agree that if women learn to perform this examen correctly, they could find a tumor earlier, when it is still very small

and the chances of avoiding a mastectomy, and may reduce the need for chemotherapy. Women should be taught by a physician how to perform this exam, but it is still not a substitute for other screening methods such as mammography.

- ***Screening Mammography.*** A mammography is a special type of x-ray imaging used to create detailed images of the breast. It uses low dose x-ray to create a high-contrast, high-resolution film, specifically intended for breast screening in women who are asymptomatic (i.e. no sign of breast cancer). It is a very common exam; in the United States, 48 million mammograms are performed each year. This non-invasive method is very effective in finding small deposits of calcium (called microcalcifications) that may be an early sign of cancer, or precancerous change in a breast. The US Food and Drug Administration reports that mammography can find 85 to 90 percent of breast cancers in women over age 50 and can discover a lump up to two years before it can be felt, i.e., in its early stages. Once a breast abnormality is found or confirmed with a mammography, additional breast imaging tests such as sonography or biopsy may be conducted.

Screening exams are intended to detect cancerous, or suspicious, changes in the breasts. On the basis of these exams, the doctor may decide that no further tests are needed, if no suspicious elements were observed in the exam. However, in some cases further analysis have to be conducted in order to diagnose observed lesions and to confirm or reject the presence of breast cancer, being biopsy the only definitive way to determine whether cancer is present. Very often, it is recommended to have a **diagnostic mammography** test first, and, if indicated by a physician, to supplement it with another test. Breast cancer diagnosis includes the following techniques:

- ***Breast MRI.*** Magnetic resonance breast imaging (MRI, MR) has been approved by the U.S. Food and Drug Administration (FDA) since 1991 for use as a supplemental tool, in addition to mammography, to help diagnose breast cancer. Breast MRI is an excellent problem-solving technology, often used to investigate breast lesions first detected with mammography, physical exams or, in general terms, other screening tests. MRI is an excellent complement to mammography, since it performs better at imaging the augmented breast, including both the breast implant itself and the breast tissue surrounding the implant, taking into account that signs of breast cancer can sometimes be obscured by the implant on a mammogram. Besides, MRI is very useful for detecting the stage of breast cancer, determining the most appropriate treatment and for patient follow-up after breast cancer treatment. Additionally, the technological features of MRI motivated researchers to analyze whether it may be useful in screening younger woman at high risk. The American Cancer Society recommends for women at very high risk of developing breast cancer to have annual breast MRI exams as a complement

- ***Ultrasonography.*** Also known as sonography or breast ultrasound, this test uses high-frequency sound waves, which can show whether a lump is a fluid-filled cyst (not cancer) or a solid mass (which may or may not be cancer). It is frequently used to evaluate breast abnormalities that are found with screening or diagnostic mammography or with a physical breast examen performed by a specialist; once again, it is recommended for this exam to be used along with mammography. This test allows significant freedom in obtaining images of the breast from any orientation and it provides high contrast resolution, but is lacks the spatial resolution (detail) of conventional mammography and, therefore, ultrasound is not approved by the United States Food and Drug Administration (FDA) as a screening tool for breast cancer. Instead, ultrasound is used to investigate a lesion detected by mammography or physical breast exam. This test is unable to image microcalcifications, which are important signs of the presence of breast cancer, while mammography is excellent at imaging them.

- ***Ductal Lavage.*** This is an investigational technique used to collect cells from milk ducts in the breast, where 95% of breast cancer starts,so that the cells can be checked for cancer under a microscope. The concept for the procedure was based on research indicating that breast cancer begins in the lining of the milk ducts, involving a series of molecular changes from normal to abnormal, and eventually to malignant. As long as abnormal cells are contained within the ducts or lobules, they are called *preinvasive*, but once they invade surrounding tissues, they are considered *invasive breast cancer*. It can take up to eight to 10 years before these cells grow into a tumor large enough to be detected by mammography (approximately one centimeter in diameter), or felt during a physical breast exam. The term *lavage* is french and means *wash* or *rinse*. Ductal lavage involves analyzing cells washed out from the breast ducts whether they have malignant qualities before they develop into breast cancer. It is a minimally invasive procedure that can be done in a doctor's office or outpatient clinic and takes about an hour. To obtain the cells, the doctor applies an anesthetic cream to the nipple and then uses a suction device to draw fluid from the milk ducts to the nipple's surface. At this point, a hair-sized catheter is inserted into the nipple, followed by a small amount of anesthetic and then a small amount of saltwater into the duct. Fluid containing cells from the duct is withdrawn through the catheter and is then removed, along with breast cells. The cells are checked under a microscope to identify changes that may indicate cancer or changes that may increase the risk for breast cancer. This test is not intended to replace mammography or a clinical breast exam, but it can help women at high risk for breast cancer assess their risk. Experts warn that this examen may not be practical for any but very-high-risk women, since it may detect many cellular abnormalities that never become cancerous.

- ***Nuclear medicine.*** Also know as *scintimammography*, this nuclear medicine breast imaging is a supplemental breast exam that may be used in some patients to investigate a breast abnormality. A nuclear medicine test is not a primary investigative tool for breast cancer but can be helpful in selected cases after diagnostic mammography has been performed. This test is performed by injecting a radioactive tracer, called *dye*, into the patient, and since it accumulates differently in cancerous and non-cancerous tissued, scintimammography can help physicians determine whether cancer is present. However, this test is not a screening tool for breast cancer. It is appropriate for those patients who have dense breast tissue that makes their mammograms difficult to interpret or in patients with palpable abnormalities but whose mammograms do not reveal any abnormalities; in any case, nuclear medicine is performed *after* a physical breast exam, mammography or ultrasound.

- ***Thermal imaging.*** Also known as *thermography* or *infrared imaging*, this test was approved by the U.S. Food and Drug Administration (FDA) in 1982 as a supplement to mammography in helping to detect breast cancer, but it is not approved as a stand-alone screening test for breast cancer. This is a way of diagnosing breast cancer by measuring and mapping the heat from the breast with the use of a special camera. A computer looks for *hot spots* or differences in heat, then analyzes the images. The theory is that an area of increased heat may indicate an increase in blood vessel formation due to cancer. However, studies have not proved this to be an effective screening tool for early diagnosis of breast cancer and it is not a replacement for mammograms; it is not a reliable diagnostic test, since it can miss some cancers and can have a high false-positive rate.

- ***Diagnostic Mammography.*** Diagnostic mammography is an X-ray exam of the breasts that is performed in order to evaluate a breast lesion or abnormality detected by physical exam or

routine screening mammography. Diagnostic mammography is different from screening mammography in that additional views of the breast are usually taken, as opposed to two views typically taken with screening mammography, hence, it is more time-demanding and costly than screening mammography. The objective of this test is to pinpoint the exact size and location of breast abnormality and to image the surrounding tissue and lymph nodes. In many cases, diagnostic mammography will help to show that the abnormality is highly likely to be benign. It this is the case, the radiologist may recommend for the patient to return at a later date for a follow-up mammogram, typically in six months. On the other hand, if an abnormality observed with diagnostic mammography is suspicious, additional breast imaging (e.g. ultrasound, MRI or others) or even a biopsy may be ordered. Biopsy is the only definitive way to determine whether a woman has breast cancer or not.

- ***Biopsy.*** It is the surgical removal of a small piece of tissue or a small tumor for microscopic examination to determine whether cancer cells are present. If a woman or her doctor finds a suspicious breast lump, or if imaging studies show a suspicious area, the woman must have a biopsy, which is the most important procedure in diagnosing cancer. While physical breast exam, mammography, ultrasound and other breast imaging methods can help detect a breast abnormality, biopsy followed by pathological (microscopic) analysis is the only definitive way to determine if cancer is present. It is estimated that over 48 million mammograms are performed each year and that less than one million of them (less than 5%) are recalled to undergo a biopsy (in some instances the number of cases requiring biopsy can be as low as 2%, depending on population demographics and methods of care). Fortunately, 65% to 80% of breast biopsies result in benign diagnosis, but if cancer is actually found to be present after pathological analysis, it is critical that the type and stage of the cancer be identified as soon as possible, in order to plan the following treatment steps. There are two types of biopsies: *needle biopsy* and *open biopsy*. The particular method chosen depends on the nature and location of the abnormality and on the patient's general health and preference. Each type of biopsy has advantages and disadvantages. There are several types of ***needle biopsies***, and they are often used first, since they are fast and simple. In a needle biopsy, the area involved is numbed and cleaned, and a sterile hollow needle is inserted through the skin to take a tissue or cell sample. The needle biopsies types include:

  1. ***Fine needle aspiration (FND).*** The doctor uses a thin, hollow needle to remove a few cells from the breast lump. It can be done in an outpatient setting and takes only a few minutes. Fine needle aspiration can also be used to remove fluid from a *cyst*. It is a percutaneous, *through the skin*, procedure and the cellular material taken from the breast is usually sent to the pathology laboratory for analysis; if the radiologist or surgeon just drains fluid from a cyst and does not send the sample to the pathology laboratory for analysis, the procedure is simply called a *cyst aspiration*
  2. ***Core needle biopsy.*** A thicker hollow needle, than the one used in FNA, is used to remove a larger amount of tissue. The skin is nicked with a scalpel so the needle can enter. This type of biopsy is done in an outpatient setting; this exam provides more tissue than fine needle aspiration and estrogen receptors can be obtained wit this method.
  3. ***Vacuum-assisted biopsy.*** This biopsy is also a percutaneous procedure that may rely on the guidance of stereotactic mammography or ultrasound imaging. A vacuum-assisted breast biopsy is done with a local anesthetic in an outpatient setting. The *mammotome* breast biopsy system is a type of vacuum-assisted biopsy that was approved in 1996; the handheld version of the *mammotome* received clearance from the U.S. Food and Drug Administration in September 1999. In this method, a large needle is inserted into the

suspicious area by using ultrasound or stereotactic guidance. The *mammotome* is then used gently to vacuum tissue from the suspicious area; additional tissue samples can be obtained by rotating the needle. This procedure can be performed with the patient lying on her stomach on a table, and if the handheld device is used, the patient may lie on her back or in a seated position. The mammogram-directed technique is called *stereotactic needle biopsy*. In this procedure, computerized mammogram breast images help the radiologist map the exact location of the breast lump and guide the tip of the needle to the right spot. The choice between a mammogram-directed stereotactic needle biopsy and ultrasound-guided biopsy depends on the type of breast change and the experience and preference of the doctor.

As stated before, there is another type of breast cancer biopsy:

1. ***Open (Surgical) Biopsy.*** In a surgical biopsy, a sample of tissue can be cut directly from the tumor that has been exposed with a surgical incision. The surgeon usually removes the entire tumor together with a margin of normal-looking breast tissue surrounding the malignant area. If the tumor cannot be felt, needle localization is done before the biopsy. After numbing the area with a local anesthetic, the surgeon places a small, hollow needle in the abnormal spot in the breast. A thin wire is inserted through the center of the needle, the needle is removed and the wire is used to guide the surgeon the right spot. A surgical biopsy may be either incisional or excisional. In an incisional surgical biopsy, only a portion of the lump is removed. It is most often performed on women who have advanced stage cancer whose tumor is too large to be removed by excisional biopsy. Excisional biopsies remove the entire lump plus some surrounding normal tissue. This is the most common type of open biopsy. Although the primary purpose is to diagnose cancer, a biopsy can algo be a surgical treatment to remove cancer.

In most cases, if it is possible, a needle biopsy is preferred to an open biopsy as the first step in a cancer diagnosis because it provides a quicker diagnosis and causes less discomfort. It also gives the woman an opportunity to discuss treatment option with her doctor before any surgery is performed. However, in some instances, an open surgical biopsy must be done instead. The surgeon can perform a core needle biopsy of fine needle aspiration if the lump can be felt; but often these procedures are done by a radiologist using ultrasound or mammogram to guide the needle, as the biopsy is of a visible, but not palpable, change. Finally, the time required of a biopsy varies according to the specific type of biopsy procedure; open biopsies that require general surgery can take much longer than others.

2.5 Breast Cancer Treatments

After been accurately diagnosed, breast cancer may be treated with local or bodywide therapy. Some patients have both kinds of treatment. Local therapy, such as surgery and radiation therapy, is used to remove or destroy breast cancer in a specific area and when breast cancer has spread to other parts of the body, such as the lung or bone, local therapy with radiation may be used to control cancer in those specific areas.

Systemic treatments are used to destroy or control cancer throughout the body. Chemotherapy, hormonal therapy, and biological therapy are systemic treatments. Some patients have systemic therapy to shrink the tumor before local therapy is performed. Others have systemic therapy to prevent the cancer from recurring or to treat cancer that has spread.

There are several breast cancer treatments, such as the following [94]:

- ***Breast-conserving therapy.*** This process is also known as breast-sparing surgery, and it is aimed at removing the cancer, but not the breast. After this kind of surgery most women receive radiation therapy to destroy cancer cells that remain in the area. *Lumpectomy* and *segmental mastectomy* (also called partial mastectomy) are types of breast-sparing surgery; both procedures are described as follows:

  1. ***Lumpectomy.*** In a lumpectomy, the surgeon removes the breast cancer and some normal tissue around it. Often, some of the lymph nodes under the arm are removed and sometimes an excisional biopsy serves as a lumpectomy.
  2. ***Segmental mastectomy.*** The surgeon removes the cancer and a large area of normal breast tissue around it. Occasionally, some of the lining over the chest muscles below the tumor is removed as well.

- ***Mastectomy.*** It is a procedure to remove the breast, or as much of the breast as possible. *Breast reconstruction* is often an option, performed either at the same time as the mastectomy or in a later surgery. Women considering reconstruction should discuss this with a plastic surgeon before having a mastectomy. *Simple* (or total) mastectomy is the removal of the whole breast without removal of lymph node. While in a *modified radical mastectomy*, the whole breast, most of the lymph nodes under the arm and often the lining over the chest muscles are removed. The smaller of the two chest muscles is also taken out to help in removing the lymph nodes. Moreover, *radical mastectomy* is the removal of the breast as well as the surrounding lymph nodes, muscles, fatty tissue and skin. Formerly considered the standard for women with breast cancer, it is rarely used today. In rare cases, radical mastectomy may be suggested if the cancer has spread to the chest muscles.

- ***Axillary lymph node dissection.*** In most cases, the surgeon also removes lymph nodes under the arm to help determine whether cancer cells have entered the lymphatic system. This is called an axillary lymph node dissection. Removing many of the lymph nodes under the arm slows the flow of lymph; that slowing may cause fluid buildup in the arm and hand, causing swelling (*lymphedema*). To prevent this, women need to protect the arm and hand on the treated side from injury or pressure, even years after surgery. A *sentinel-lymph node biopsy* is offered at some cancer centers. Researchers are hoping that this procedure may reduce the number of lymph nodes that must be removed during breast cancer surgery. After injecting a radioactive substance that flows through the lymphatic system to the first lymph nodes where cancer cells are likely to have spread (the sentinel nodes), the surgeon makes a small incision and removes only the nodes that have radioactive substance or blue dye. Then, a pathologist checks the sentinel lymph nodes for cancer cells; if no cancer cells are detected, removing additional nodes may not be necessary.

- ***Radiation therapy.*** Women who have had a lumpectomy usually receive radiation therapy after the surgical wound has healed to kill any remaining cancer cells. The radiation may be directed at the breast by an external machine or may occur from radioactive material in thin plastic tubes that are placed directly into the breast (*brachytherapy*). Some women have both kinds of radiation therapy. Radiation therapy is sometimes also used before surgery, to destroy cancer cells and shrink tumors. It may be given alone or with chemotherapy or hormonal therapy. This approach is most often used in cases in which the breast tumor is large or not easily removed by surgery.

- ***Systematic treatment.*** It is a kind of treatment of the body to prevent cancer recurrence. *Chemotherapy* is the use of drugs to kill cancer cells; it is usually given by stage IV and when used after surgery it is used for a definite period of time (three to six months). Side effects are determined by which drugs are used.

- ***Hormonal therapy.*** This treatment is intended to block the estrogen needed by some cancers to grow. Its side effects depend on the kind of drug or treatment. The drug *tamoxifen* is the most common hormonal treatment; it blocks the cancer cells' use of estrogen but does not stop estrogen production. Tamoxifen may cause hot flashes, vaginal discharge or irritation, nausea and irregular periods. Women who are still menstruating may become pregnant more easily when taking tamoxifen.

- ***Biological therapy.*** It is a treatment designed to enhance the body's natural defenses against cancer. For instance, *trastuzumab* (Herceptin) is a monoclonal antibody that targets breast cancer cells that have too much of a protein known as *human epidermal growth factor receptor-2* (HER-2). By blocking HER-2, Herceptin slows or stops the growth of these cells. Herceptin may be administered alone or with chemotherapy. This is used in women whose cancers are HER-2-neu positive and have spread to other organs. The side effects of biological therapy depend on the types of substances used. Rashes or swelling at the injection site are common, and flu-like symptoms also may occur. Herceptin may cause these and other side effects, but these effects generally become less severe after the first treatment.

There are several factors that a woman's treatment options depend on. These factors include her age and menopausal status, her general health, the size and location of the tumor and the stage of the cancer, the results of lab tests, and the size of her breast.

Woman who have early stage breast cancer (stages 0 through II) may have breast-sparing surgery followed by radiation therapy to the breast, or they may have a mastectomy, with or without breast reconstruction. Sometimes radiation therapy is also given after mastectomy. Breast-sparing surgery and mastectomy are equally effective; the choice depends mostly on the size and location of the tumor, the size of the woman's breast, certain features of the cancer and the woman's preference about preserving her breast.

Patients who have stage III breast cancer usually have both local treatment to remove or destroy the cancer in the breast and chemotherapy or hormonal therapy to prevent the disease from spreading. On the other hand, women who have stage IV breast cancer receive chemotherapy and/or hormonal therapy to destroy cancer cells and control the disease.

2.6 Summary

In this chapter we described the medical fundamentals of breast cancer, including factors that increase the chances of developing that disease, the related symptoms that would determine the presence of breast cancer in potential patients and the different types of breast cancer, being *ductal carcinoma in situ* the most frequent one, accounting for 90% of non-invasive cases. Also, we described the different stages in which this cancer evolves, from *stage 0*, a non-invasive carcinoma, to *stage IV*, a metastatic cancer; a patient under a *recurrent stage* is that who has been diagnosed positive, followed the medical treatment, and developed cancer nonetheless.

Moreover, we described the different techniques and technologies that are used for breast cancer screening, being mammography the most popular one, since it reveals important breast features to the radiologist that may be a sign of an *early* stage of this disease. However, there are related exams

that can be performed, including breast MRI, ultrasound and even breast self-exams in which patients themselves explore the breast area in the search for palpable tumors. Naturally, biopsy is another alternative and possesses the highest accuracy for diagnosis this disease.

If a person is diagnosed as having breast cancer, there are different medical treatments that can be followed; all of them were described in this chapter. It is important to recall that there exists effective medical treatment to cure this disease as long as it is detected in the earliest stages. Otherwise, complications may arise and radical medical response may be necessary, such as performing a *mastectomy*, which is a procedure to remove the breast, either partially or completely.

Chapter 3

Mammograms

This chapter will provide information about mammograms, as a tool for early breast cancer diagnosis, considering all the characteristics that make them a standard in screening this disease in early stages. The screening process itself is also depicted along with the different lesions and the features of them that can be evidence of breast cancer and that mammography can reveal.

A *conventional* mammography is a special type of X-ray imaging that uses low dose X-ray to create detailed images of the breast, in a high-contrast, high-resolution film and relies on an X-ray system designed specifically for imaging breasts.

A screening mammography is the best tool available for finding breast cancer early, before symptoms appear. Screening mammograms are used to look for breast changes in women who have no signs of breast cancer. And, as it has already been discussed in previous sections, successful treatment of breast cancer depends on early diagnosis, in which mammography plays a major role, given that it provides the radiologist with relevant features of the breast that can be analyzed in order to detect and diagnose cancer, in its earliest stages.

Mammography can often detect breast cancer before it can be felt and can reveal small deposits of calcium in the breast, which represent an important evidence for the presence of the disease. Although most calcium deposits are benign, a cluster or very tiny specks of calcium, called *microcalcifications* may be an early sign of cancer.

More than 90 percent of all breast cancers are detected by a mammogram [94]. Most doctors agree with the current American Cancer Society recommendation of screening mammograms every year for women at age 40. However, this means that 10 percent of breast cancers are missed; it is more common to miss cancers in women with dense breasts, which is a condition seen more often in younger women.

A typical mammography screening includes two views of each breast (see section 3.3): one from above, and one from the side. Normally, the technician examines the X-ray pictures immediately to make sure pictures are clear. A physician then looks at the mammograms and if a mass, changes, from earlier mammograms, abnormalities of the skin, or enlargement of the lymph nodes is found, further testing may be recommended. This could include an ultrasound of the breast, a biopsy or needle sampling, or consultation with a breast surgeon.

Results of between 5 percent and 10 percent of mammograms are abnormal. Of those in younger women that are followed up with additional tests (another mammogram, fine-needle aspiration, ultrasound, MRI or biopsy) most are not cancer.

The following sections provide a description for the screening mammography process, including a description of the digital type of a mammogram, the process of taking a mammogram itself, the different findings that the radiologist may observe in a mammogram, and the features that early breast

cancer presents in a mammogram.

3.1 Digital Mammography

One of the most recent advances in X-ray mammography is digital mammography. Digital (computerized) mammography is similar to standard mammography in that X-rays are used to produce detailed images of the breast. Digital mammography uses essentially the same mammography system as conventional mammography, but the system is equipped with a digital receptor and a computer instead of a film cassette. Several studies have demonstrated that digital mammography is at least as accurate as standard mammography [64].

In standard mammography, images are recorded on film using an X-ray cassette. The film is viewed by the radiologist using a *light box* and then stored in a jacket in the facility's archives. As for digital mammography, the breast image is captured using a special electronic X-ray detector, which converts the image into a digital picture for review on a computer monitor; the digital mammogram is then stored on a computer. With digital mammography, the magnification, orientation, brightness, and contrast of the image may be altered after the exam is completed to help the radiologist more clearly see certain areas.

Digital mammography systems cost approximately 1.5 to 4 times as much as standard film mammography systems. However, procedural time saved by using digital mammography over standard film mammography may justify part of the cost for facilities that perform several thousand mammograms each year.

From the patient's perspective, a digital mammogram is the same as a standard film-based mammogram in that breast compression and radiation are necessary to create clear images of the breast. The time needed to position the patient is the same for each method. However, conventional film mammography requires several minutes to develop the film while digital mammography provides the image on the computer monitor in less than a minute after the exposure/data acquisition. Thus, digital mammography provides a shorter exam for the woman and may possibly allow mammography facilities to conduct more mammograms in a day. Digital mammography can also be manipulated to correct for under or over exposure after the exam is completed, eliminating the need for some women to undergo repeat mammograms before leaving the facility (i.e. the radiologist may see certain ares of the breast more clearly, since the magnification, orientation, brightness and contrast of the mammogram image may also be altered after the exam is completed).

Is it believed that in the near future, digital mammography may provide many benefits over standard film mammography. These benefits include:

- Improved contrast between dense and non-dense breast tissue.
- Faster image acquisition (less than a minute).
- Shorter exam time (approximately half that of film-based mammography).
- Easier image storage.
- Physician manipulation of breast images for more accurate detection of breast cancer.
- Ability to correct under or over-exposure of films without having to repeat mammograms.
- Transmittal of images over phone lines or a network for remote consultation with other physicians.

Preliminary results of the Digital Mammographic Screening Trial (DMIST) [79], show that digital mammography may be more accurate at detecting breast cancer in some women than standard film mammography. According to the study results, digital and standard film mammography had similar accuracy rates for many women. However, digital mammography was significantly better at screening women in any of the following categories:

- Women under age 50, regardless of what level of breast tissue density they had.
- Women of any age with very dense or extremely dense breasts
- Women in pre- or perimenopausal stage of any age (defined as women who had a last menstrual period within 12 months of their mammograms).

The study showed no benefit for post-menopausal women over age 50 who did not have dense breast tissue.

While digital mammography is quite promising, it still has additional hurdles to undergo before it replaces conventional mammography. Digital mammography must:

- Provide higher detail resolution (as standard mammography does).
- Become less expensive (digital mammography is currently several times more costly than conventional mammography).
- Provide a method to efficiently compare digital mammogram images with existing mammography films on computer monitors.

Standard mammography using film cassettes has the benefit of providing very high detail resolution (image sharpness), which is especially useful for imaging microcalcifications (tiny calcium deposits) and very small abnormalities that may indicate early breast cancer. While full-field digital mammography may lack the spatial resolution of film, clinical trials have shown digital mammography to be at least equivalent to standard film screening mammography. This is because digital mammography has the benefit of providing improved contrast resolution, which may make abnormalities easier to see. Various manufacturers are trying to develop digital mammography systems with detail resolution equivalent to standard film mammography while also providing the benefits of digital mammography noted above.

Finally, the high cost of digital mammography is a major obstacle and, additionally, standard mammography systems are currently installed in over 10,000 locations across the United States; it may take years for this current equipment to be updated or replaced and for digital mammography to become widespread.

3.2 Screening Mammography Process

According to [69, 65], during mammography, the technologist will position the patient and image each breast separately. One at a time, each breast is carefully positioned on a special detector plate and then gently compressed with a paddle (often made of clear Plexiglas or other plastic). This compression flattens the breast so that the maximum amount of tissue can be imaged and examined.

Mammography technologists may place adhesive markers to the breast skin prior to taking images of the breast. The purpose of the adhesive markers is twofold: first, to identify areas with moles, blemishes or scars so that they are not mistaken for abnormalities, and secondly, to identify areas that

may be of concern (for instance, a lump was felt during physical examination). Some centers routinely mark the nipple with a small dot to provide a clear "landmark" for the radiologist on the mammogram images.

To take a mammogram, the low dose X-ray source is turned on and X-rays are radiated through the compressed breast and onto a digital sensor located within the shelf that provides support under the breast. The X-rays penetrate the digital sensor to produce either a high-contrast, high-resolution film, or a detailed and electronic image of the internal structures of the breast. For digital mammography, highly sensitive digital detectors and special X-rays are used for mammography to create the highest quality images at the lowest exposure.

It is the special energy and wavelength of the X-rays that allow them to pass through the breast and create the image of the internal structures of the breast. As the X-rays pass through the breast, they are attenuated (weakened) by the different tissue densities they encounter. Fibrous breasts are very dense and absorb or attenuate a great deal of the X-rays. The connective tissue around the breast ducts and fat is less dense and attenuates or absorbs far less X-ray energy. It is these differences in absorption and the corresponding varying degrees of exposure to the film (for traditional mammography) or digital detector (for digital mammography) that create the images which can clearly show normal structures such as fat, fibroglandular tissue, breast ducts, and nipples. Further, abnormalities such as microcalcifications (tiny calcium deposits), masses, and cysts are also visible.

If a digital mammography is performed, the digital image appears on a computer screen within seconds after the exam which allows the technologist to ensure the best possible images of the breast were obtained. The images are then processed through a sophisticated computer analysis system called CAD (computer-aided detection) which serves as a computerized second opinion in the review of your images. The images are then interpreted by a radiologist, who compares the new images of a woman's breast to each other and to previous mammograms a woman has had. The radiologist will look for shadows and patterns of tissue density to detect any abnormalities.

A mammogram is like a fingerprint; the appearance of the breast on a mammogram varies tremendously from woman to woman, and no two mammograms are alike. It is extremely helpful for the radiologist to have films (not just the report) available from previous examinations for comparison purposes. This will help the doctor to recognize small changes that occur gradually over time and detect a cancer as early as possible. For this reason, it is important that a woman carefully considers where her mammograms are performed and establishes a mammographic history with an accredited mammography facility.

The breast is made of fat, fibrous tissue and glands. Breast masses (these include benign and cancerous lesions) appear as white regions on mammogram images. Fat appears as black regions on a mammogram image. Everything else (glands, connective tissue, tumors and other significant abnormalities such as microcalcifications) appear as levels of white on a mammogram.

3.3 Views of Screening Mammography

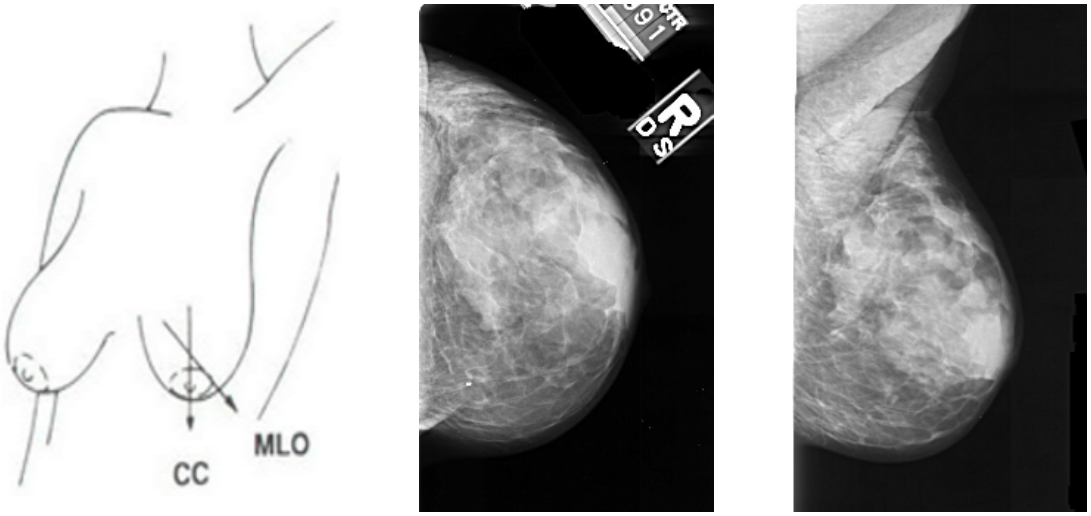
As previously mentioned, a typical screening mammography process involves taking two views of each breast, one from above, called *cranial-caudal* view (CC) and one from an oblique or angled view, called *mediolateral-oblique* view (MLO) the side.

Moreover, diagnostic mammography may involve taking supplemental view tailored to the specific problem. These may include views from each side: *latero-medial* view (LM) which is taken from the side towards the center of the chest, and *mediolateral* view (ML), which is taken from the center of the chest out. It may also include exaggerated cranial-caudal and other special views such as spot compression and magnification views.

The views taken during the screening mammography test are described in the following [66, 70]:

- ***Mediolateral Oblique (MLO).*** It is taken from an oblique or angled view. During routine screening mammography, the MLO view is preferred aver a lateral 90-degree projection because more of the breast tissue can be imaged in the upper outer quadrant of the breast and the axilla. Figure 3.1(a) depicts the direction of this view, and Figure 3.1(c) shows the resulting mammography (taken from *The Digital Database for Screening Mammography (DDSM)*). With the MLO view, the pectoral muscle (upper left corner in Figure 3.1(c)) should be depicted obliquely from above and visible down to the level of the nipple or further down. The shape of the muscle should curve or bulge outward as a sign that the muscle is relaxed; the medial (middle) portion of the breast should be prominent in the MLO view. It is important that compression be applied over the whole image area. The nipple should be depicted in profile and a small stomach fold should be visible as a sign that the whole breast is reproduced.

- ***Cranio-caudal view (CC).*** The cranio-caudal view images the breast from above. This view may be taken during routine screening mammography and during diagnostic mammography. Figure 3.1(a) depicts the direction of the projection and Figure 3.1(b) shows its corresponding mammography. With the CC view, the entire breast parenchyma (glandular tissue) should be depicted. The fatty tissue closest to the breast muscle should appear as a dark strip on the X-ray and behind that it should be possible to make out the pectoral muscle. The nipple should also be depicted in profile.



(a) CC and MLO views.

(b) Mammogram showing a CC view.

(c) Mammogram showing a MLO view.

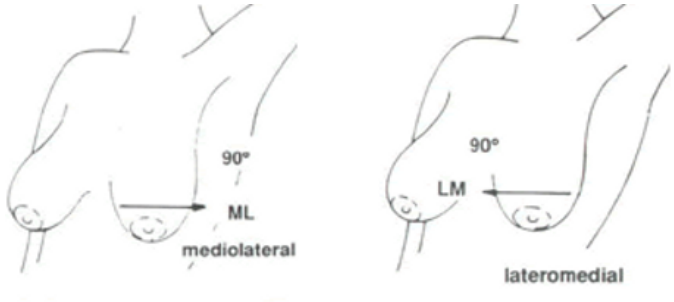
Figure 3.1: The different views taken in screening mammography, depicting the direction of the view (a), and the corresponding mammographies (b) (taken from the DDSM, file A_0030_1.RIGHT_CC) and (c) (taken from the DDSM, file A_0030_1.RIGHT_MLO)

Furthermore, supplemental views taken during diagnostic mammography are:

- ***Medio-lateral view (ML).*** This view is taken from the center of the chest outward. If no oblique projection is taken, the mediolateral position may be preferable to the latero-medial view, since the lateral side of the breast, where pathological changes are most commonly found, is then

closest to the film. However, if the physician wants to include as much of the medial side of the breast as possible, the lateromedial view may be chosen. Figure 3.2(a) depicts the direction of this view. With a lateral view, the pectoral muscle should be depicted as a narrow light band on at least half of the picture. the nipple should be depicted in profile and a clear stomach fold should be visible under the breast.

- ***Latero-medial view (LM).*** This view images the breast from its outer side toward the center of the chest. When physicians want to include as much of the medial portion of the breast, the LM view may be used. Figure 3.2(b) depicts the direction when taking this view.



(a) Medio-lateral view (ML). (b) Latero-medial view (LM).

Figure 3.2: The medio-lateral view (a) and latero-medial view (b), depicting the directions of the projection.

3.4 Features Observed in Mammograms

There are several small abnormalities or changes that are of interest for a radiologist to perform a breast cancer analysis, since they are an evidence for the presence of the disease. A classification of the common findings that can be observed in a mammogram is provided in the following [94]:

1. ***Normal.*** Interpretations of mammograms can be difficult because a normal breast can appear differently for each woman; therefore, there is not a general description about the appearance of a normal mammogram. It is more practical to define the abnormal findings.

2. ***Calcifications.*** There are two types of calcifications that may be observed in a mammogram. *Macrocalcifications* are large calcium deposits that appear in the breast as a result of aging, old injuries or inflammation. These deposits are related to noncancerous conditions and do not require a biopsy. Macrocalcifications occur in about half of all women older than age 50 and in one of 10 women younger than age 50, in the United States.

   Microcalcifications are tiny specks of calcium in the breast that may appear alone or in clusters. The shape and location of microcalcifications can help a radiologist determine how likely it is that the areas are malignant. In some cases, microcalcifications do not require a biopsy, but only a follow-up mammogram within three to six months. In other cases, the microcalcifications are suspicious and a biopsy is recommended.

3. ***Masses.*** A mass may occur with or without calcifications and can be caused by benign breast condition or by breast cancer. Some masses can be monitored with periodic mammography; others may need a biopsy. The size, shape and margins of the mass help the radiologist to determine the likelihood of cancer. Many masses turn out to be *cysts* (benign collections of fluid); to confirm that a mass is really a cyst, a doctor must either order a breast ultrasound or remove some fluid with a needle.

   If a mass is not a cyst, then the patient may need more imaging tests. Prior mammograms may help show that a mass has remained unchanged for many years, indicating a benign condition. If a mass raises a significant suspicion of cancer, tissue must be removed for examination under the microscope to determine whether it is cancer. This can be done with needle biopsy or open surgical biopsy.

4. ***Architectural distortions.*** They are the result of a demoplasic reaction where a focal interruption of the normal tissue pattern occurs, and the surrounding tissue slightly distortions towards a focal point. The lesion itself is not visible, but it can be detected by analyzing the surrounding tissues.

5. ***Asymmetric density.*** They describe density variations in equivalent regions of both breasts. It is not expected to find the exact same features in both breasts, but this analysis is aimed to detect structures that could seem to be normal fatty tissues when only one breast is examined, but may become suspicious when they are absent in the equivalent region of the other breast.

Furthermore, the American College of Radiology has developed a standard way of describing mammogram findings by giving the results a code numbered 0 through 5, called the *Breast Imaging Reporting and Data System* (BI-RADS) [62]. The BI-RADS is a guide to breast cancer diagnostic routines, and radiologists sometime refer to each category as being a level:

1. ***Category 0.*** Assessment is incomplete and additional imaging evaluation is needed. A possible abnormality may not be completely seen or defined and requires additional evaluation including the use of spot compression, magnification views, special mammographic views or ultrasound.

2. ***Category 1.*** No significant abnormality to report. The breasts are symmetrical without masses, distortion or suspicious calcifications.

3. ***Category 2.*** This is a negative mammogram result that has found a benign lesion, such as benign calcifications, intra-mammary lymph nodes, and calcified fibroadenomas. This category ensures that other individuals viewing the mammogram do not misinterpret a benign finding as suspicious, and documents the finding for use in future mammogram assessments.

4. ***Category 3.*** This is a *probably benign finding*, which suggests the need for a short-term follow-up. Results in this category are not expected to change. However, since the results have not been proved benign, the doctor will want to see whether the lesion changes over time. In this case, follow-up imaging is usually done every six months for a year, and then every year for two years. This schedule eliminates unnecessary biopsies but ensures that any malignancy will be detected within a short period.

5. ***Category 4.*** This result is a suspicious abnormality, requiring a biopsy. In this case, although the findings do not definitely appear to be cancer, there is a substantial probability of malignancy.

6. ***Category 5.*** These findings are characteristic of cancers, with a high probability of malignancy. Biopsy is very strongly recommended.

A summary for the BI-RADS categories is provided in Table 3.1.

BI-RADS Categories	
Category 0	Need additional Imaging Evaluation.
Category 1	Negative.
Category 2	Benign finding.
Category 3	Probably benign finding: short interval follow-up suggested.
Category 4	Suspicious abnormality: biopsy should be considered.
Category 5	Highly suggestive of malignancy: appropriate action should be taken.

Table 3.1: BI-RADS Assessment Categories

3.5 Mammographic Features of Early Breast Cancer

This section will provide a description of those mammographic features that are present at the earliest stages of breast cancer, when they are non-palpable. It should be recalled that the study of early breast cancer features is of high importance, since it will lead to an accurate diagnosis in a stage when the disease can still be cured.

The greatest advantage of mammography is its ability to detect breast cancers before they grow large enough to be palpable. Mammographic detection of breast cancer at the earliest possible stage requires optimal radiographic technique and a full knowledge of the subtle features with which very small cancers can present. Although some early cancers are identified as characteristic clusters of calcifications or as spiculated or multi-nodular masses, other demonstrate less typical and sometimes much less obvious mammographic signs: the single dilated duct, focal architectural distortion, asymmetry, and the developing density sign. These features are broadly classified into two categories: ***conventional*** and ***indirect*** signs, which will be described in the following [85]:

- **Conventional Signs.** There are two conventional signs that are evidence for the presence of breast cancer:

    1. ***Clustered Calcifications.*** Many non-palpable breast cancers are detected by the mammographic demonstration of clustered microcalcifications, which are smaller than 0.5 mm; they are so small that the radiologist must search for them systematically with the aid of a magnifying lens. The smallest of mammographically detectable cancers are found this way. Most experts agree that at least five such calcifications within a 1 cm^3 volume define a *cluster*, although some would broaden the definition to encompass as few as three calcific particles. Figure 3.3 shows an example of clustered calcifications.

        Some mammographers have described diagnostic criteria to differentiate benign from malignant clustered calcifications, on the basis of shape, with thin linear, curvilinear and branching shapes suggesting malignancy, and, on the other hand, round or oval shapes indicating benign lesions. However, use of these criteria has proved difficult, probably because some calcific particles are so small that it is impossible to resolve their individual shapes with conventional mammographic techniques; the use of magnification mammography results in most calcifications being classified properly.

    2. ***Poorly defined mass.*** The second characteristic mammographic presentation of malignancy is a poorly defined mass of irregular contour. These masses, even when containing

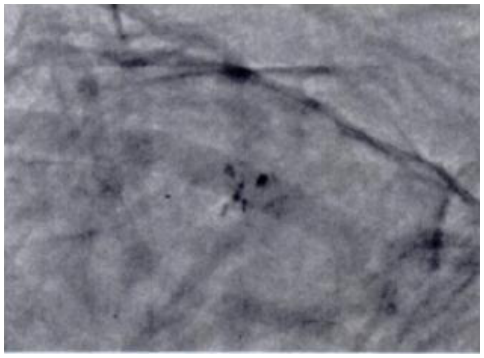


Figure 3.3: Clustered calcifications.

no calcifications, are readily imaged in older women because the fatty tissues surrounding them provide sufficient contrast to permit detection.

Small deposits of benign fibroglandular tissue often can be distinguished by the scalped, concave contours with which they usually present. However, there are a sizable number of small benign masses whose margins also appear to be poorly defined, and therefore are difficult to differentiate from malignancy, resulting in the need to biopsy several benign lesions in order to remove each early cancer. Once again, magnification mammography can be very helpful in correctly characterizing masses. Figure 3.4 provides an example of a mass with poorly defined margins; many of these lesions prove to be benign, but some are malignant.

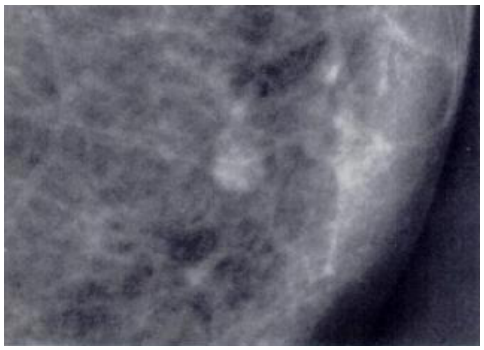


Figure 3.4: Mass with poorly defined margins.

- **Indirect signs.** Despite the best mammographic technique, it is inevitable that there will be small malignancies that demonstrate neither masses nor calcifications. Some of these may be detected by recognizing one of several mammographic signs that indicate the presence of early cancer only **indirectly**. Indirect signs of breast cancer include the following:

  1. ***Single Dilated Duct.*** Mammographic demonstration of dilation of a single retroareolar duct can be the only indication of an underlying intraductal carcinoma, since the cancer and the duct itself almost always are non-palpable, and since these lesions often do not cause nipple discharge that is sufficiently worrisome for the patient to seek medical attention.

The mammographic appearance of a dilated duct is characteristic, that of a tubular and branching structure, widest in caliber at the nipple, tapering as it proceeds distant into the parenchyma. Some mammographers consider the demonstration of more than one adjacent dilated duct to have similar significance to that of a single dilated duct, but these broadened criteria are not universally accepted. In either case, the underlying cause much more often is benign than malignant, except in the unusual circumstance when the duct contains clustered calcifications characteristic of carcinoma, as depicted in Figure 3.5.

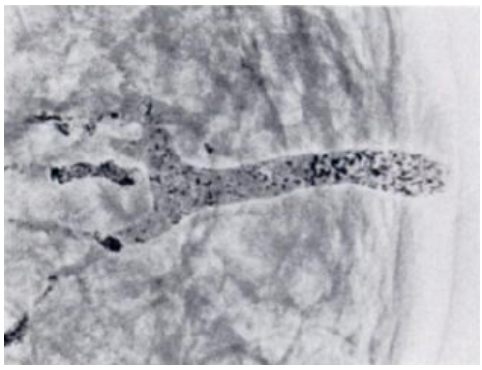


Figure 3.5: Dilated duct with calcifications.

2. ***Architectural distortions.*** Invasive carcinoma distorts the interfaces between fat and normal breast parenchyma due to the desmoplastic response of host tissues. In the fatty breast, there is no diagnostic dilemma, since this results in the typical spiculated mass seen by mammography. However, in the very dense breast, the tumor mass can be so obscured by adjacent benign tissues as to be invisible, leaving as the only indication of underlying malignancy an area of focal architectural distortion. Depending on the location of the cancer, the retractive phenomena producing architectural distortion will result in different mammographic appearances. Figure 3.6 shows an example of an architectural distortion, with few of its spiculations marked by the arrows.

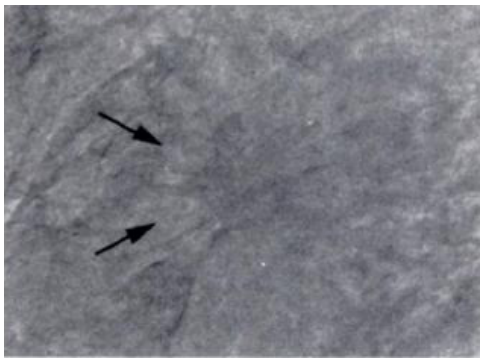


Figure 3.6: Architectural distortion showing few of its spiculations.

3. ***Asymmetry.*** When an early carcinoma appears on mammography as a poorly defined mass without knobby margins or radiating spiculations, it still can be detected readily in a fatty breast. However, it is difficult to identify a poorly defined malignant mass as being significantly different from the several other poorly defined densities elsewhere in

the breast. One helpful perceptual aid in this situation is to view the mammograms of both breasts side by side, projection for projection, to facilitate recognition of a small cancer by the asymmetry with which it appears. Figure 3.7 depicts the analysis of this asymmetric issues, by observing two bilateral projection mammograms.

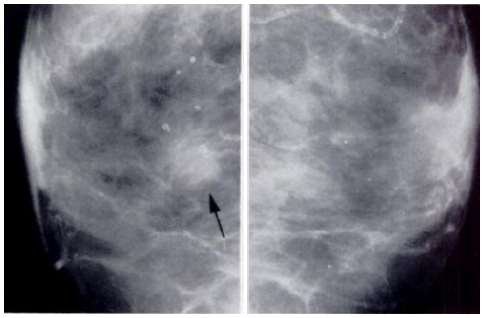


Figure 3.7: Asymmetry analysis; bilateral lateral projection mammograms.

4. ***Developing density.*** An early cancer may demonstrate mammographic features so subtle that it shows neither a poorly defined mass, nor focal architectural distortion, nor asymmetry. The only clue to its existence may be the *interval appearance* on mammograms of small focus of increased density. It should be taken into account that, except during puberty, pregnancy, lactation and periods of exogenous estrogen stimulation, the breast is an involuting organ whose natural history involves progressive fatty replacement; therefore, it is reasonable to consider as potentially malignant any newly apparent focus of increased mammographic density. Moreover, the developing density sign is nonspecific, because benign lesions develop and grow just as cancers do; nonetheless, it is an important sign because it can be the **earliest** indicator of breast cancer.

   To identify a developing density, one must have previous mammograms available for comparison. Therefore, it is mandatory to locate and obtain prior mammograms before completing the interpretation of current examinations. Figure 3.8 depicts the process of observing if a patient has developed density in the breast, by analyzing previous and recent mammograms.

Figure 3.8: Observing developing density by analyzing previous and recent (from left to right) mammograms.

3.6 Summary

In this chapter we described mammographies as an effective and popular tool for *early* breast cancer screening and diagnosis. A mammography is a special type of imaging that uses a low dose X-ray to

create detailed images of breast tissue with which more than 90% of all breast cancers are detected. It can be used in women who do not have any symptoms of the disease, as a prevention mechanism. A digital mammography, as opposed to the conventional type, presents some advantages because several screening features can be modified to help the radiologist to see regions of interest more clearly, including the magnification, orientation, brightness, and contrast of the image, although the equipment for digital mammography is more expensive.

Also, we presented the process of screening breast cancer using mammograms, including a description of how the four different views are taken in every new case, with the objective of collecting information from different perspectives of the breast. Additionally, we described the features that physicians look for in a mammogram and the related lesions that may be evidence of the presence of this disease, being the microcalcification clusters and masses the two most common ones.

Chapter 4

Image Mining

This chapter explains the relevant theoretical background on *Image Mining*, including *Content-based Image Retrieval* which will be used in this research work as a means of improving a classifiers performance.

Image mining is useful for gathering information from an image database, based on a set of *features* that may be text-specified by the user. However, this process would result in the need of *indexing* and labeling all images in the database, which could be a tremendous effort if we consider a large database. *Content-based Image Retrieval* alleviates this problem by automatically extracting features from a query image and retrieving those images similar to it from the database; thus, there is also the need to specify a means to measure the similarity between images. In order to perform this *content-based* retrieval, there are several types of features that can be taken into account to generate a so-called *image signature*, which is a n*-dimensional* vector that describes an image.

However, image mining is a difficult task, since, unlike structured datatypes, the image datatype is not suitable for interpretation by a machine; the content of an image is visual in nature and the interpretation of the information in it depends on the given observer (subjective). Moreover, data mining techniques have been developed mainly for structured datatypes, such as text, which is somewhat *easier* to mine.

Extracting features from an image and its interpretation seems to be an effortless task for the human visual system, but modeling the human interpretation of the semantic content of images is a major research effort and, thus, it is difficult to come up with a set of algorithms that can serve as *complete* image mining tools.

There is a debate about what exactly is image mining and how it differs from image processing, since the latter has also been used to detect patterns, retrieve images by content and pattern matching; there is no closed answer, but while image processing focuses on detecting abnormal patterns and retrieving images, image mining deals with making associations between different images from large image databases [91].

Processes in mining an image database have been focused on the search and retrieval of images based on the analysis of similarity of a query image or its features with the entries in the image database. Furthermore, retrieval systems are broadly categorized based on the type of searches in [72]:

1. ***Using a description of an image.*** Images are described based on user-defined texts, which are used for both image indexing and retrieval; they can include size, type, date of capture or some text description of the image. This process is called *text-based image retrieval*. All descriptions provided are typed manually by human operators.

2. ***Using visual content.*** This process is called *Content-Based Image Retrieval*, where images are

retrieved based on their visual content; image features are extracted and used for indexing the database to enable the searching process of *similar* images.

As one can see, the text-based descriptions cannot be automated without considering a visual feature extraction provided by a human and, hence, this process is very time-demanding and almost impractical for the large amount of image databases available nowadays; besides, descriptions provided are subjective and therefore they cannot be automated accurately. Consequently, *content-based image retrieval* is highly desirable, since the mining process is based on an automated visual feature extraction.

4.1 Image Retrieval

An image retrieval system is essentially a database management system for handling image data, which can be x-rays, pictures, satellite images, and photographs. It is necessary an accurate data model for image representation and a functional architecture for the overall system, that would include a *query manager*, *browser*, *editor*, *update manager*, *storage manager* and *metadata manager*, as well as an *integrity and security manager*, as depicted in Figure 4.1 [91].

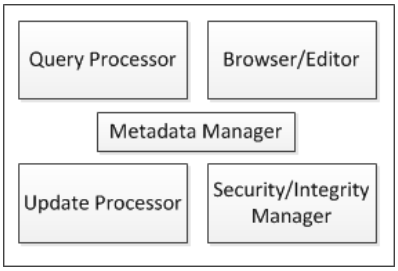


Figure 4.1: Architecture for an image retrieval system.

Users may interact with the image retrieval system in different ways. One way to measure user-system interaction level is the complexity of queries, which are supported by the system; from a user's perspective, they can be regarded as the different *query modalities* or options available for the user to query the system. Those modalities include [26]:

- ***Keywords.*** This is the typical search in which the user provides the system with words (text) that are used to perform a query.
- ***Free-text.*** The user provides a complex sentence, phrase, question or story as a query. This input describes the images that the user wants to retrieve from the database.
- ***Image.*** When the user tries to retrieve images similar to a *query image*, that is provided as an input. If no reliable metadata is available, using an example image is the most representative way to query a system, to retrieve images based on their content.
- ***Graphics.*** The user provides a hand-drawn or computer-generated picture, as a query.
- ***Composite.*** When two or more of the aforementioned modalities are used to query a system, including interactive querying schemes such as *relevance feedback* systems.

The previous query modalities require for the system to have the appropriate methods to support user-interaction. As one may expect, these processing methods become more complex when visual queries and/or user interactions are involved. From a system perspective, query processing methods include:

- ***Text-Based.*** This type of query processing usually is regarded as performing one or more simple keyword-based searches and then retrieving matching pictures. Some form of natural language processing may also be involved, since understanding the query as a whole may be expected.
- ***Content-Based.*** Content-based query processing involves the design of an automated process to extract visual features from the query image, in order to form an *image signature*, and then to retrieve all images similar to the query image from the database, by computing a *similarity measure* between images; this automated process, based on visual content is formally called *content-based image retrieval* (CBIR).
- ***Composite.*** This type of processing may include both content- and text-based processing methods in varying proportions.
- ***Interactive-Simple.*** When the system must support the interaction of users, considering only one query modality (*either* text or images). Relevance-feedback-based image retrieval systems are examples of this type of processing method.
- ***Interactive-Composite.*** User interaction to the system is supported with more than one modality: using *both* text and images, for instance. This is the most advanced form of query processing that is required to be performed by an image retrieval system.

It is important to note that the text-based methods require a human observer to provide visual information, in order to type manually a description text or a tag text, which is typically subjective to the observer, hence, making more difficult to develop an automated scheme; this process is known as *text-based image retrieval*. The *content-based image retrieval* (CBIR) paradigm, on the other hand, is based on the *visual content* of the image, which is extracted in an automated way without the need of any subjective observers, and therefore it is a more practical and accurate method. CBIR will be described in the following section.

4.2 Content-based Image Retrieval

As mentioned above, a Content-based Image Retrieval (CBIR) system is the process to extract images from a database, which are similar to a query image, based on their visual content. CBIR, as we see it today, is any technology that in principle helps to organize digital picture archives by their visual content; taking this into account, anything ranging from an image similarity function to a robust image annotation engine falls under CBIR. This is why it is a field of study located at a unique juncture within the scientific community, where people from different fields such as computer vision, machine learning, information retrieval, human-computer interaction, database systems, Web and data mining, information theory, statistics and even psychology are contributing as part of the CBIR community.

By the nature of its task, CBIR systems, in general terms, involve two intrinsic problems [26]:

1. ***How to mathematically describe an image.*** This is important for retrieving images from a database, based on their *visual content*, because the original representation of an image, which is an array of pixel values, barely corresponds to our human visual response, and semantic

understanding of the image. For retrieval purposes, the mathematical description of an image is referred to as its *signature*, which is used to design similarity measures between images.

2. ***How to assess the similarity between a pair of images.*** Based on their abstracted descriptions, or *signatures*, a similarity measure is computed to know the extent to which the visual content of two given images is similar.

These issues, the extraction of *signatures* and the calculation of *image similarity*, cannot be cleanly separated, since the formulation of signatures determines to a large extend the realm for definitions of similarity measures. In terms of methodology development, a strong trend which has emerged in recent years is the use of statistical and machine learning techniques in several aspects of the CBIR technology. For instance, automatic learning, mainly clustering and classification, is used to form either *fixed* or *adaptive* signatures, to tune similarity measure, and even to serve as the technical core of certain searching schemes, like *relevance feedback*, in which the user interacts with the system in the searching process.

There are, in general, three fundamental modules in a CBIR system, according to Sushmita et al [72]:

1. Visual content or feature extraction,
2. Multidimensional indexing, and
3. Retrieval.

The architecture for a possible CBIR system is depicted in Figure 4.2. We can see that this architecture is mainly divided into two modules. The first one processes all images from the image database in an off-line manner; in this process, the features from each image in the image database are extracted to form the *meta-data* information of the image, in order to describe the image using its visual content features. Next, these features are used to index the image, and then they are stored into the meta-data database along with the images. The second part of the process describes the retrieval task; in this stage the query image is analyzed to extract visual features, which are used to retrieve the similar images from the image database. Then, instead of directly comparing two images, similarity of the visual features of the query image is measured with the features of each image stored in the meta-data database as their signatures; often, the similarity of images are measured by computing the distance between the feature vectors of the two images. The process results in the system returning the first k images whose distance from the query images is below some defined threshold.

Images in a database are indexed based on extracted inherent visual contents, or *features* such as color, texture, shape, topology, salient points, etc. Then, an image can be represented by a multidimensional vector of the extracted features, which actually forms its signature. The most popular of them are color, texture, shape and salient points in an image [26], which will be described in the following sections.

4.3 Feature Extraction

As depicted in Figure 4.2, it is necessary to perform a *feature extraction* process in all images from the image database. Feature extraction is a means of extracting compact but semantically valuable information from images, which can be represented by a multidimensional vector of the extracted features; this information is referred to as the *image signature*, which will be used to index the images and build the *image meta-data* database. This vector can be assumed to be associated to a point in

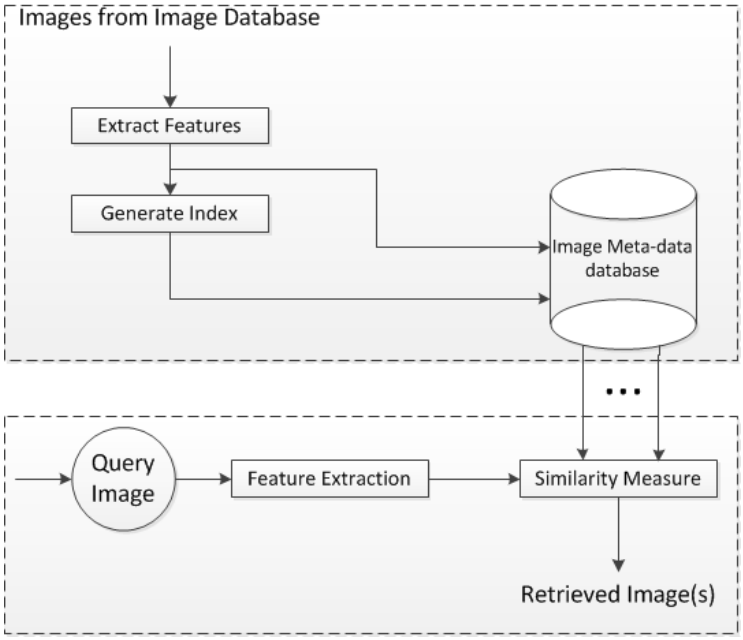


Figure 4.2: Architecture of a CBIR system [72].

the multidimensional space. For instance, an image can be represented by an N-dimensional feature vector whose first n_1 components may represent *color*, the next n_2 components may represent *shape*, the following n_3 components may represent some image *topology*, and finally n_4 components may represent *texture* of the image, so that there are $N = n_1 + n_2 + n_3 + n_4$ components [72].

This way, an *example image* can simply be used as a query using visual-content-based indexing; in this process, the query image will be analyzed to extract its visual features, in order to be compared to find matches with the indices of the images stored in the meta-data database. The feature vectors of similar images will then be clustered in the N-dimensional search space, and the process of retrieving images *similar* to the query will be focused on finding the indices of those images whose feature vectors are within some threshold of proximity to the point represented by the vector of the query image.

Extractable features are often divided into:

1. ***Primary features.*** Refers to the raw image data.
2. ***Secondary features.*** The features derived from the raw image data.
   (a) *Primitive features.* Features computed from the raw image data.
   (b) *Logical features.* The identity or name of an object in the image.
   (c) *Abstract features.* Refers to the environment in which the image was made or the cause why it was made.

This research work will consider ***primitive features***, for the *image retrieval* processes, since the extraction of them is relatively simple and fast, and, thus, they are often used. In fact, the proposed model uses the actual feature vectors that are extracted from the segmented lesions, after they are passed through a feature selection process.

Furthermore, there are two options when extracting features from an image [26]:

1. ***Global extraction.*** Features are computed to capture the overall characteristics of an image (as a whole). The advantage of global extraction is its high speed for both extracting features and computing similarity. However, global features are often too rigid to represent an image, as they can be oversensitive to location and thus fail to identify important visual characteristics

2. ***Local extraction.*** It is used to obtain a region-based signature; prior to the feature extraction, a reliable *image segmentation* process has to be performed. Local extraction is typically followed by an extra step of feature summarization or clustering. A set of features are computed for every pixel using its neighborhood (a block around the pixel). Another option is for the image to be divided into small, non-overlapping blocks, and then to compute features from every one of them; this option reduces computation and the features are still local because of the small block size.

Section 4.9 will provide further details on how to build a so-called *image signature*, based on the extraction of features.

Furthermore, features aimed at identifying an object in an image should have the following four characteristics [17]:

1. ***Discrimination.*** It means that features represent objects in different classes, which should have significantly different values.

2. ***Reliability.*** Refers to features that describe the objects in the same class, which should have similar values.

3. ***Independence.*** Feature should not be strongly correlated to each other. This does not mean that the larger the number of the features, the better they represent an object.

4. ***Small number.*** Some redundant features can even make the decision worse; it is a problem of dimensionality. Hence, a small number of features that can accurately represent the object in its different classes or shapes is desirable.

In the next sections a description of typically used features is provided, as well as the *image segmentation* process, which is a step that is often performed prior to the actual extraction of features.

4.4 Feature Selection

The motivation for applying feature selection (FS) techniques has become a prerequisite for model building in many applications. In particular, the high dimensional nature of many modelling tasks in different areas, such as classification and pattern recognition, has given rise to a wide variety of feature selection techniques.

In contrast to other dimensionality reduction techniques like those based on projection (e.g. Principal Component Analysis) or compression (e.g. using information theory), feature selection techniques do not alter the original representation of the variables, but merely select a subset of them, hence, preserving the original semantics of the variables and offering the advantage of being interpreted by a domain expert.

The most important objectives of performing feature selection are [82]:

1. To avoid overfitting and improve model performance or, more specific, prediction performance in the case of supervised learning and better cluster detection in the case of unsupervised learning.

2. To provide faster and more cost-effective models.

3. To gain a deeper insight into the underlying processes that generated the data.

However, searching for a subset of relevant, highly discriminative features, introduces an additional layer of complexity in the modelling task, and, of course, an additional computational effort. In the classification context, feature selection techniques can be organized in two categories, depending on how they combine the feature selection search with the construction of the classification model. Those categories are:

- ***Filter methods.*** The relevance of features is computed by looking at the intrinsic properties of the data. In this model a score for feature relevance is calculated, and those features with a low score are removed. In a subsequent stage, this subset of features is presented as input to the classification algorithm. Some **advantages** of filter techniques are: they easily scale to very high-dimensional datasets, they are computationally simple and fast, they are independent of the classification algorithm. Therefore, this approach has to be performed only once, and, afterwards, different classifiers can be evaluated, using the previously computed subset of features. On the other hand, a disadvantage of filter methods is that they ignore the interaction with the classifier and that most techniques are univariate, which means that each feature is considered separately, hence, ignoring dependencies among features; this issue can lead to a low performance when compared to other types of techniques. However, to alleviate this problem, several multivariate filter techniques were introduced, with the objective of incorporating dependencies between features to some degree. Examples of this technique include: χ^2, information gain, $i-$test, etc., regarding univariate models, and Correlation-based feature selection (CFS), Markov blanket filet (MBF), among others, for multivariate techniques.

- ***Wrapper methods.*** In this approach, the classification algorithm is embedded within the feature subset search. A search procedure in the space of possible feature subsets is defined, and various subsets of features are generated and evaluated; the evaluation of a given subset of features is computed by training and testing a classification method. In order to search the space of all feature subsets, a search algorithm is then *wrapped* around the classification model. As the space of feature subsets grows exponentially with the number of features, heuristic search methods are used to guide the search for an optimal subset. Moreover, these search methods can be divided in two classes: *deterministic* and *randomized* search algorithms. Some **advantages** of wrapper approaches are: the interaction between feature subset search and model selection and the ability to take into account feature dependencies; on the other hand, a common drawback is that they have a higher risk of overfitting than filter techniques and are computationally intensive, especially when building the classifier has a high computational cost. Some examples of this approach are: Sequential Forward Selection (SFS), Sequential Backward Elimination (SBE) or Beam Search, among other *deterministic* search algorithms, and Simulated Annealing, Randomized Hill Climbing, Genetic Algorithms, etc., for *randomized* search.

The feature selection approach that was studied and implemented in this research proposal belong to the *wrapper* category, as a *genetic algorithm* was designed to randomly search for a subset of highly discriminative features, considering (separately) all four classification methods that are of interest for this research project, which are: k-NN classifier, Neural Network (NN), Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA). During the development of the proposed model, we found out that the classification performance that we achieve with wrapper feature selection techniques is much better than the one observed with filter methods; this is, we empirically determine to use a wrapper approach.

4.5 Image Segmentation

Most CBIR systems perform feature extraction as a preprocessing step [26]. The obtained visual features serve as inputs to subsequent image analysis tasks, such as similarity estimation or annotation. In order to extract region-based visual signatures (only those present in important parts of an image), a *segmentation* process has to be conducted as a first step.

Image Segmentation takes a complete image as an input, and identifies those *regions of interest* (ROIs), which are the critical portions of an image that interest to the researcher or observer. For instance, reliable segmentation is especially critical for characterizing shapes within images, without which the shape estimates are largely meaningless.

There are several research efforts focused on designing an accurate segmentation phase, in order to extract ROIs, given a specific scenario or problem. For instance, segmentation based on k-means clustering approach is widely used, due to its speed. However, some other recent methods have been developed, all of which have advantages and disadvantages, in addition to the fact that some approaches can be more appropriate than others at solving a specific segmentation problem; thus, a complete analysis of the scenario's characteristics should determine the best segmentation technique to be used. See section 6.1, for a detailed reference on related image segmentation techniques, applied to digital mammograms.

Once this process is complete, features from ROIs are extracted and analyzed in order to use them for the CBIR process; commonly used features will be described in the following sections.

4.6 Color features

In spite of the fact that this research work is not going to consider *color features*, since the mammograms are gray-level images, this section will provide a description of this topic, since *color* is one of the visual features most widely used in CBIR.

While we can perceive only a limited number of gray levels, our eyes are able to distinguish thousands of colors and a computer can represent even millions of distinguishable colors in practice. Color has been successfully applied to retrieve images, because it has very strong correlations with the underlying objects in an image. Furthermore, color as a feature is robust to background complications, scaling, orientation, perspective, and size of an image. A color pixel in a digital image is represented by three color channels, which usually are Red, Green and Blue (RGB); it is well know that any color can be produced by mixing these three primary colors.

However, other color spaces such as CIELAB and CIELUV (both created by the International Commission on Illumination, CIE in french), as well as HSV (Hue, Saturation and Value), and HLS (Hue, Lightness and Saturation), were found to deliver better results as compared to the RGB space, because these color spaces are perceptually uniform compared to the RGB and, consequently, they are more effective to measure color similarities between images.

Since it is widely used in the CBIR community, the HSV color space is described in the following:

- ***Hue.*** Represents the relative color appearance (i.e. the "redness", "greenness" and "blueness").
- ***Value.*** It indicates the darkness of the color, or perceived illuminance.
- ***Saturation.*** It represents the strength of the color.

Color features include:

1. ***Color histogram.*** It is the most commonly used color feature in CBIR. It has been found to be very effective in characterizing the **global** distribution of colors in an image, and it can be used as an important feature for image characterization. To define color histograms, the color space is quantized into a finite number of n discrete levels $r_1 \dots r_n$, that are usually equally distributed over the color space, and each of them become a bin in the histogram. The color histogram is then computed by counting the number of pixels in each of these discrete levels; that is, by finding the perceptually nearest r_i for each pixel in the image, and updating the statistics on this reference color.

2. ***Color coherence vector.*** The problem with color histogram-based similarity approach is that the global color distribution doesn't reflect the spatial distribution of the color pixels locally in the image and, thus, doesn't distinguish whether a particular color is sparsely scattered all over the image or it appears in a single large region in the image. *Color coherence* approach classifies each pixel in an image, based on whether it belongs to a large uniform region or not. For instance, a region can be considered uniformly colored if it consists of the same color and the area of the region is above a certain threshold of the whole image area; these pixels are referred to as *coherent* pixels. By distinguishing this, color coherence vectors provide superior retrieval results as compared to the global color histogram method.

3. ***Color moment.*** This is a compact representation of the color feature to characterize a color image. It has been shown that most of the color distribution information is captured by the three low-order moments: first-order moment (μ) captures the mean color, the second-order moment (σ) captures the standard deviation, and the third-order moment (θ) captures the skewness of color. These three low-order moments ($\mu_c, \sigma_c, \theta_c$) are extracted for each of the three color planes c; as a result, we need to extract only nine parameters, corresponding to three moments for each of the three color planes, to characterize the color image. Finally, the weighted euclidean distance between color moments of two images, has been found to be effective to calculate color similarity.

4. ***Linguistic color tag.*** Global color distribution using color histogram does not take advantage of the fact that the adjacent histogram bins might actually represent roughly the same color, because of the limited ability of the human perceptual system. Moreover, only a limited number of color shades are sufficient for visual discrimination between two images. To take advantage of this, a color matching technique based on *linguistic tags*, focused on identifying a color with a name was proposed. The concept behind this technique is to construct *equivalence classes* of colors, which are identified by *linguistic tags*, or color names such as pink, red, etcetera, that perceptually appear the same to the human eye but are distinctly different from that of neighboring subspaces. Thus, the dimensionality of color features is significantly reduced and computations for color similarity measures decrease as well.

We can see that when we use histograms, either *color histograms* or *color coherence vectors*, we often will have to deal with very high dimensional vectors, in order for us to get an accurate description of the content of an image. Moreover, similarity measures for histograms are sometimes complicated, as the similarity of every two reference colors has to be considered. Consequently, sometimes the *color moments* are preferred. It all depends on the specific problem or scenario, since it will define the type of features that best fits its requirements.

4.7 Texture features

In spite of the fact that there is no standard or formal definition for *texture* in the literature, one major characteristic of it is the *repetition of pattern or patterns over a region in an image*. The elements of those patterns are sometimes called *textons*, which can vary in size, shape, color and orientation over the region. Moreover, the degree of variation of the textons is a measure of the difference of two textures, as well as their spatial distribution int he image.

Texture is an innate property of virtually all surfaces, such as bricks, clouds, skin, etc. It contains information regarding underlying structural arrangement of the surfaces in an image. When a small area in an image has wide variation of discrete tonal features, the dominant property of that area is *texture*; on the other hand, the *gray tone* is a dominant property when a small area in the image has very small variation of discrete tonal features.

Several techniques have been used for measuring textural similarity, such as the representation of texture feature as a co-occurrence matrix, in order to mathematically represent gray level spatial dependence of texture in an image; it is constructed based on the orientation and distance between image pixels. Meaningful statistics are extracted from this co-occurrence matrix, as the representation of texture. Since basic texture patterns are governed by periodic occurrence of certain gray levels, co-occurrence of gray levels at predefined relative positions can be a reasonable measure of the presence of texture and periodicity of the patterns. Features such as *entropy*, *energy*, *contrast*, and *homogeneity*, can be extracted from the co-occurrence matrix of gray levels of an image.

Besides, popular signal processing techniques have also been used in texture analysis and extraction of visual texture features. Wavelet transforms have been applied in texture analysis and classification of images, based on multiresolution decomposition of the images and representing textures in different scales. Of all the different wavelet filters, Gabor filters were found to be very effective in texture analysis.

4.8 Shape features

Shape is another image feature applied in CBIR. It can be defined as the description of an object minus its position, orientation and size. Therefore, shape features should be invariant to *translation*, *rotation*, and *scale*, for an effective CBIR process, when the arrangement of the objects in the image are not known in advance.

Shape features are mainly used for technical images where the number of colors is small and the shape of the objects in the image is relatively well defined. To use *shape* as an image feature, it is essential to segment the image to detect object or region boundaries, and sometimes this process is not trivial.

Shape characterization can be divided in two categories:

1. ***Boundary-based.*** It uses the outer contour of the shape of an object.

   (a) *Fourier Descriptors.* This is the most prominent representative of *boundary-based* shape characterization. It uses the Fourier-transformed boundaries of the objects as the shape features.

2. ***Region-based.*** It uses the complete shape region of the object.

   (a) *Moment Invariants.* This is the most representative method of this category; it uses the region-based geometric moments that are invariant to translation and rotation. There are

seven normalized central moments that can be taken as shape features, which are also scale invariant.

4.9 Image Signature and Similarity Measure

Once visual features have been extracted, they are used to build a so-called *image signature*, which is a well defined structure that describes either the entire image or a region of it; as previously mentioned, if it is decided to use region-based visual features, an accurate *image segmentation* process is an essential step, prior to the actual extraction of features.

According to Datta [26], considering the mathematical formulations, the types of signatures can be broadly classified into *vectors* and *distributions*, considering that histograms and region-based signatures can both be regarded as sets of weighted vectors and, when the weights sum up to one these sets are equivalent to discrete distributions (discrete in the sense that the support is finite). The most exploited type is the region-based signature and, thus, it is important to mention its mathematical connection with histograms, for computation purposes.

However, it should be noted that distributions extracted from a collection of local feature vectors can be of other forms such as a continuous density function or even a spatial stochastic model. A continuous density is generally more precise in describing a collection of local feature vectors than a discrete distribution with finitely many support vectors. For special kinds of images, these sophisticated statistical models may be needed to characterize them.

In Figure 4.3, it is depicted the overall image signature formulation process [26]; it can be observed that, as mentioned in Section 4.3, for feature extraction purposes there are two options: either we extract them in a global or a local basis. If a local extraction is performed, the following process is to summarize or clusterize the set of extracted features. It can also be observed that the user may interact in some extent with the retrieval system, in order to provide feedback that may be used to adapt the formulation of the image signature to meet the user's needs.

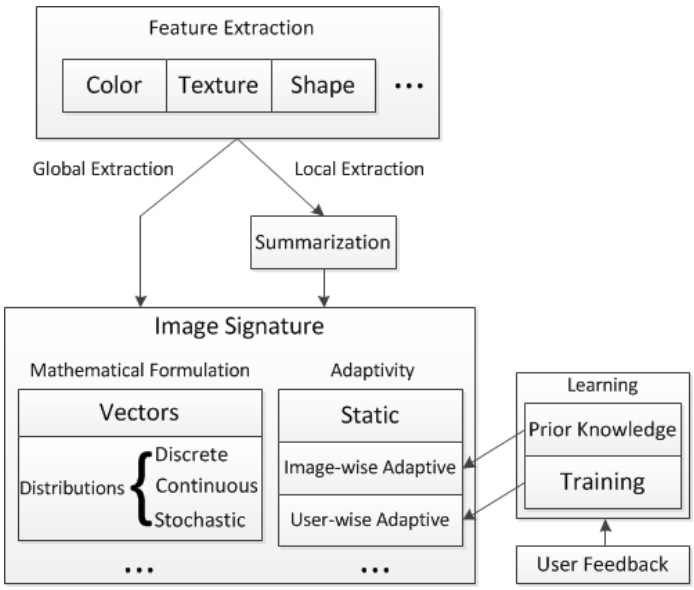


Figure 4.3: Formulation of Image Signature [26].

Once the the image signature formulation has been made, the next step is to determine a way to accurately measure the *similarity* of two given images, based on their signatures and to use this

computation in the retrieval process, in order to obtain a number of images which are the *most similar* to the query image.

We have to consider that the basis of the extraction determines the type of signature that we are going to obtain; i.e., if we decide to use a region-based signature, it will consist of a set of vectors (one for each region of the original image) and, on the other hand, a summary of local feature vectors will provide a signature in the form of a distribution. Moreover, a single vector could be used as signature for an entire image or a given region of interest within the image. This is an important consideration, since the type of signature we are working with is going to determine its mathematical formulation and, consequently, the different techniques one can use as a *similarity measure*. Figure 4.4 shows the different techniques for computing *similarity* of images considering the type of signature.

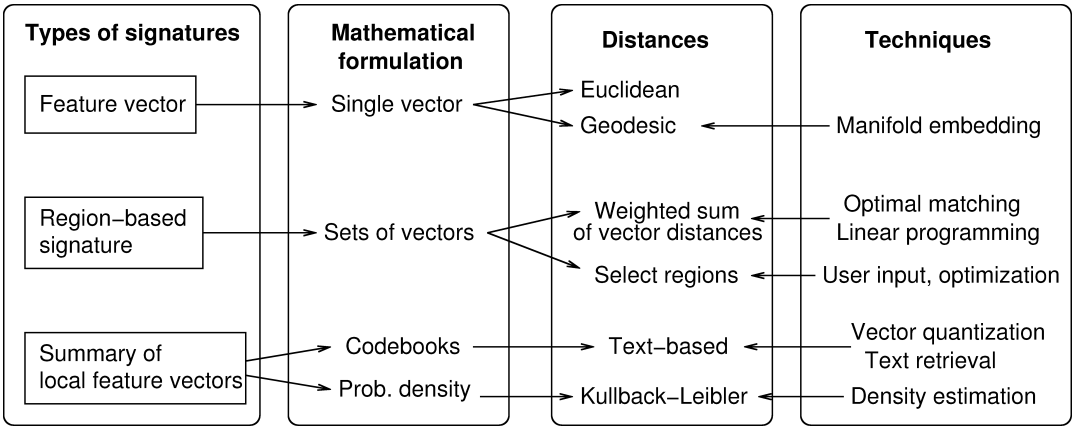


Figure 4.4: Types of Image Similarity Measure and related techniques [26].

In this research work we are going to use a vectorial image signature with continuous values, coming from the set of features that will be computed from the detected lesions. A single vector will be constructed for each region of interest, even if there are more than one of them in the same mammographic image. Therefore, in accordance to the literature review presented in this section, we are going to use a set of distance-based metrics (Euclidean, Manhattan, among others) to determine the similarity between two given feature vectors. Specific details on this matter will be provided in Chapter 7.

4.10 Summary

The solution model proposed in this research work makes use of several *image processing* and *image mining* techniques. Therefore, in this chapter we thoroughly described the fundamentals of these research topics, including the fundamentals of Content-Based Image Retrieval, a process that is useful for extracting images from a database, which are similar to a given query image, considering an appropriate representation of the visual content (or image signature) of both the query and each of the stored images.

This retrieval mechanism is computed based on an abstraction of the features that are extracted from images, such as a vectorial representation of all the components; several types of features that can be computed from images were also described. Then, unless there is a solid understanding of what attributes are more relevant for the retrieval process, feature vectors are passed through a *feature selection* stage in which the dimensionality is reduced by either implementing a *filter* or a *wrapper* approach in which they are analyzed to find the subset that provides the highest discriminant power.

To compute the similarity of two given images, there are several signatures (or representations) that can be considered to make an abstraction of their visual content and each of those determine the similarity metrics that are more appropriate to compute, being vectors of features the most common image signatures. All the related fundamentals of image similarity and how to design an image signature were introduced in this chapter as well.

Chapter 5

Machine Learning Theory

The problem of searching for data patterns has been extensively addressed by machine learning methods; its learning approach is composed of a set of known elements $\{x_1, ..., x_n\}$ called a *training set*, and a *target vector* t, which represents the category of the corresponding element. The categories of the training set are known in advance and there are different methods to represent categories in terms of vectors [9].

The result of running such an algorithm is represented by a function $y(x)$, which takes an input x and computes an output vector y. The form of the function $y(x)$ is determined during the *training* or *learning* phase of the algorithm, based on the training data. When the model is trained it is capable of determining the category to which new elements belong. These new elements are taken from a separate *test set*, and the ability to correctly categorize new elements, others than the ones contained in the training set, is called *generalization*.

The learning approaches within machine learning are divided in the following areas:

Supervised Learning: includes applications in which the training data is composed of both the input vectors, and their corresponding target vectors. Problems in which the input vectors should be assigned to a discrete number of categories are called *classification* problems; on the other hand, if we need the output to comprise one or more continuous variables, the approach is called *regression*.

Unsupervised Learning: in this type of problems, the provided training data do not include any corresponding target values, and the objective is to recognize groups of similar examples within the data (*clustering* problems) or to compute the distribution of data within the input set (*density estimation*).

Reinforcement Learning: its goal is to maximize a reward which is given for the actions taken in a given situation.

The problem that this research work addresses is a *binary classification* scheme, in which we will compute several features from breast lesions, in order to use them as inputs to the selected algorithms and to assign each new case a category or, formally, a *class*, which in this case will inform if the lesion is *malignant* or *benign*.

The next sections will provide a definition and description of the *classification* task, along with a categorization of the different methods that are commonly encountered in the literature, according to [72]; it is important to notice that this categorization may vary according to the author. Afterwards, the four classification algorithms that are of interest for the development of the solution model presented in this document will be described. Then, *Case-based Reasoning* is also described, as a form

of learning in advanced decision support systems. Finally, we also provide a description of the Fuzzy Rule Induction Algorithm (FURIA) and classifier ensembles, as well as techniques that can be used to deal with the class-imbalance problem which is typically present in bio-medical datasets and several other domains.

5.1 Classification

Also described as supervised learning, *classification* is a scheme in which there is a database of tuples, each assigned a class label, and the objective is to develop a model or profile for each class. This model is then used to predict the class $C_i = f(x_1, \ldots, x_n)$ of new cases, where $x_1, \ldots, x_n$ are the input attributes contained in the tuple database. The database will be partitioned into training and test sets; the classifier is trained on the former, and the latter is used to asses the generalization capability of the classifier, as we will further explain later in this chapter.

As stated in [72], the input to the classification algorithm is, typically, a dataset of training records with several attributes (or features) and one so-called *dependent attribute*, which states the class of the record. The remaining *predictor attributes* can be numerical or categorical in nature; a numerical attribute has continuous, quantitative values, while a categorical one takes discrete, symbolic values that can also be class labels. If the dependent attribute is categorical, the problem is called classification with this attribute being considered as the class label; on the other hand, if it is numerical, the problem is termed regression.

The goal of classification (and regression) is to build a concise model of the distribution of the dependent attribute in terms of the predictor attributes. **Decision boundaries** are generated to discriminate between patterns belonging to different classes. The resulting model is used to assign values to a database of testing records, where the values of the predictor attributes are known but the dependent attribute is to be determined.

Classification methods can be categorized as follows:

1. **Decision Trees [80].-** The decision space is divided into piecewise constant regions. Typically, an information theoretic measure is used for assessing the discriminatory power of the attributes at each level of the tree.

2. **Nearest neighbor classifiers [93].-** It is an approach for classifying data or objects, based on closest training examples in the feature space. This approach is amongst the simplest of all the machine learning algorithms, in which the distance from all instances is computed and the an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its neighbors.

3. **Regression [13].-** The goal of regression is to predict the value of one or more continuous *target* variables t given the value of a D-dimensional vector x of *input* variables. These models can be linear or polynomial, of the form $ax_1 + bx_2 + c = C_i$.

4. **Neural Networks [36].-** These incorporate learning in a *data-rich* environment, such that all information is encoded in a distributed fashion among the connection weights between a series of neurons. This model is motivated by the properties of the human brain, which is a massively parallel distributed processor, made up of simple processing units, called neurons.

5. **Probabilistic models [93].-** They are a type of *Linear Models* for classification, in which decision surfaces area linear functions of an input vector x and hence are defined by $(D-1)$-dimensional hyperplanes within the D-dimensional input space. These methods compute probabilities for hypotheses based on Bayes' theorem.

6. **Discriminant Functions.-** A **discriminant** is a function that takes an input vector x and assigns it to one of K classes, denoted C_k. There are *linear* or *non-linear* discriminant functions in the literature, which can be useful to classify data that are *linearly separable* or not, respectively.

5.2 Neural Networks

Work on artificial neural networks, typically referred to as simply *neural networks*, has been motivated by the fact that the human brain performs in an entirely different way from the conventional digital computer. The brain is a *highly complex*, *nonlinear* and *parallel computer*. In the general form, a *neural network* is a machine that is designed to *model* the way in which the brain performs a particular task or function of interest.

Formally, a neural network is a massively parallel distributed processor made up of simple processing units, which has a natural propensity for storing experimental knowledge and making it available for use. By making an analogy of the central nervous system and the neuron elements such as the axons, dendrites ans synapses, the neural networks resemble the brain in two ways [36]: (1) knowledge is acquired by the network from its environment through a learning process, and (2) the inter-neuron connection strengths, known as synaptic weights, are use to store the acquired knowledge.

The process by which the network modifies the synaptic weights is called a *learning algorithm*, through which it attains a desired design objective in an orderly fashion. Moreover, it is also possible for a neural network to modify its own topology, which is motivated by the fact that neurons in the human brain can die and that new synaptic connections can grow.

The computing power of the network is derived from its massively parallel distributed structure and also its ability to learn and therefore generalize; in this context, *generalization* refers to the neural network producing reasonable outputs for inputs not encountered during training (learning). These two information-processing capabilities make it possible for neural networks to solve complex (large-scale) problems that are currently intractable.

Properties that the use of neural networks bring, include:

1. **Nonlinearity.** Artificial neurons can be linear or nonlinear. A neural network made up of an interconnection of nonlinear neurons, is itself nonlinear; additionally, this nonlinearity is of a special kind, since it is *distributed* throughout the network. Having the property of nonlinearity es very important, specially if the underlying mechanism which generates the inputs is inherently nonlinear.

2. **Input-Output Mapping.** The *supervised learning* or *learning with a teacher* paradigm involves the modification of the synaptic weights of a neural network by applying a set of labeled *training samples*, each of them consisting of an *input signal* and a corresponding *desired response*. This way, the network is presented with an example picked at random from the set and the synaptic weights of the network are modified to minimize the difference between the actual response and the desired one. Thus, the network learns from the examples by constructing an *input-output mapping* for the problem.

3. **Adaptivity.** Neural networks have a built-in capability to *adapt* their synaptic weights to changes in the surrounding environment. This way, networks can deal with problems in which the environmental conditions change.

4. **Evidential Response.** In pattern classification, networks can be designed to provide information not only about which particular pattern to select, but also about the confidence of the decision.

5. **Contextual Information.** Knowledge is represented by the very structure and activation state of a neural network; every neuron in the network is potentially affected by the global activity of all other neurons in the network, thus, handling contextual information in a natural way.

6. **Fault Tolerance.** A neural network has the potential to be fault-tolerant, or capable of robust performance, in the sense its performance degrades gradually under adverse operating conditions.

7. **VLSI Implementability.** The massively parallel nature of a neural network makes it potentially fast for the computation of certain tasks. This same feature makes a neural network well suited for implementation using *very-large-scale-integrated* (VLSI) technology.

8. **Uniformity of Analysis and Design.** Neural networks are universal as information processors, in the way that same notation is used in all domains which involve the application of neural networks. Moreover, neurons are *common* to all types of neural networks, making it possible to share theories and learning algorithms.

9. **Neurobiological Analogy.** The design of a neural network is motivated by analogy with the brain, which is a living proof that fault tolerant parallel processing is not only physically possible but also fast and powerful. Neurobiologists look to artificial neural networks as a research tool for the interpretation of neurobiological phenomena. In a similar way, engineers look to neurobiology for new ideas to solve problems more complex than those based on conventional hard-wired design techniques.

According to the manner in which the neurons of a neural network are structured, we may identify three network architectures:

1. **Single-Layer Feedforward Networks.** Networks which have an input layer of source nodes that projects onto an output layer of neurons (computation nodes) but not viceversa. This means that the network is strictly *feedforward* or *acyclic*; we do not count the input layer of source nodes because no computation is performed there.

2. **Multilayer Feedforward Networks.** They have one or more layers of hidden neurons that are inaccessible from both the input and output sides of the network. The function of the hidden neurons is to intervene between the external input and the network output in some useful manner; by adding one or more hidden layers, the network is enabled to extract higher-order statistics.

3. **Recurrent Networks.** This type of networks has at least one *feedback* loop, in which neurons in the output layer feed its response back to the input neurons of the network (global feedback). Also, neurons could have *self-feedback* loops (local feedback). The application of feedback provides the basis for short-term memory, and provides a powerful basis for the design of nonlinear dynamical models.

This research project will only consider the *multilayer feedforward* type, since the nature of the problem that is going to be addressed requires for the classifier to consider the nonlinearity that can be obtained by adding a hidden layer. Figure 5.1 shows an example of this network topology; in this case, network is considered to be *fully connected* since every node in each layer is connected to every other node in the adjacent forward layer. If some communication links are missing, it is said to be *partially connected*.

As previously mentioned, learning is accomplished by using examples and to feed them to the network, for it to adjust its internal parameters (weights) and, thus, be able to classify new instances

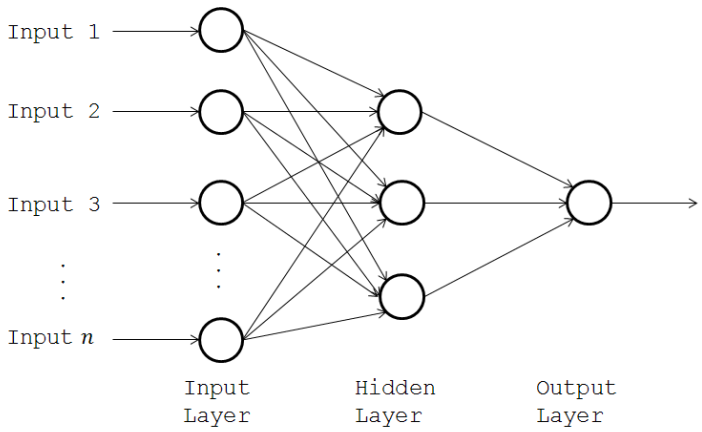


Figure 5.1: Architecture of a Multilayer Feedforward Neural Network

in the testing phase; this process is also called *training*. There are three learning paradigms to consider when designing a neural network which relate to the three learning categories mentioned before, meaning that neural networks can be used in all three schemes:

1. **Supervised.** This paradigm is based on direct comparison between the actual output of the network and the desired correct output (target). It is often formulated as the minimization of an error function such as the total Mean Square Error (MSE) between the actual output and the desired output, summed over all available data. A gradient descent-based optimization algorithm such as *Backpropagation* [43] can then be used to adjust connection weights in the network iteratively, in order to minimize the error.

2. **Reinforcement.** Reinforcement learning is a special case of supervised learning where the exact desired output is unknown. It is based only on the information of whether or not the actual output is correct. The learning of an input-output mapping is performed through continued interaction with the environment in order to minimize a scalar index of performance.

3. **Unsupervised or self-organized.** Unsupervised learning is solely based on the correlations among input data. No information on *correct output* is available for learning. Rather, it is designed a *task-independent measure* of the quality of representation that the network is required to learn, and the free parameters of the network are optimized with respect to that measure.

Finally, there are several *learning rules* that are of major importance for each of the aforementioned *learning algorithms*. These are basically weight-updating rules which determine how connection weights are changed. There are five well-known learning rules: error-correction learning, memory-based learning, Hebbian learning, competitive learning and Boltzman learning. The application of each of them depends on the properties of the learning problem.

5.3 Support Vector Machines

The Support Vector Machine (SVM) [95] has recently become very popular as a high-performance classifier in several domains, including pattern recognition, data mining and Bioinformatics. It has strong theoretical foundations and a good generalization capability. Another advantage of the SVM is that, as a by-product of learning, it obtains a set of support vectors, which characterizes a given

classification task or compresses a labeled dataset. Often, the number of support vectors is only a small fraction of the original dataset.

The objective is to construct a hyperplane as the decision surface, such that the **margin** of separation between positive and negative examples is **maximized**. The *structural risk minimization* principle is used for the purpose. Here, the error rate of a learning machine on test data (generalization error rate) is considered to be bounded by the sum of the training error and a term depending on the Vapnik-Chervonenkis (VC) dimension.

VC dimension is a measure of the *capacity*, or complexity, of a statistical classification method, which is defined as the cardinality of the largest set of points, or data samples, that the algorithm can separate, for all assignments of labels to those points and **with no errors**. Stated in another way, it is the maximum number of points that can be separated in all possible ways, by a given classifier.

In the nonlinear SVM, the discriminant hyperplane, given a labeled set of N training samples $(\mathbf{X}_i, y_i)$, where $\mathbf{X}_i \in R^n$ and $y_i \in -1, 1$, is defined as:

$$f(\mathbf{X}_q) = \sum_{i=1}^{N} y_i \alpha_i K(\mathbf{X}_q, \mathbf{X}_i) + b \tag{5.1}$$

In equation (5.1), $K(\cdot)$ is a kernel function and the sign of $f(\mathbf{X}_q)$ determines the membership of the query sample $\mathbf{X}_q$. Kernel functions are used to make (implicit) nonlinear feature map, and the first kernels investigated for pattern recognition problems are:

$$K(\mathbf{x}, \mathbf{x}_i) = (\mathbf{x} \cdot \mathbf{x}_i + 1)^d \tag{5.2}$$

$$K(\mathbf{x}, \mathbf{x}_i) = e^{\|\mathbf{x}-\mathbf{x}_i\|^2/2\sigma^2} \tag{5.3}$$

$$K(\mathbf{x}, \mathbf{x}_i) = \tanh(\kappa \mathbf{x} \cdot \mathbf{x}_i - \delta) \tag{5.4}$$

The kernel function in equation (5.2) results in a classifier that is a **polynomial** of degree d in the data. On the other hand, equation (5.3) gives a Gaussian **Radial Basis Function** (RBF) classifier, while equation (5.4) results in a particular kind of two-layer **sigmoidal hyperbolic tangent** neural network.

As it was previously stated, the structural risk has to be minimized in order to find the optimal separating hyperplane (the one which the largest margin of separation between positive and negative samples); this leads to the need of solving a quadratic programming (QP), which may involve a dense $N \times N$ matrix, where N is the number of points in the dataset. Since most QP routines have quadratic complexity, SVM design requires huge memory and computational time for large data applications, which is a limitation of the model [72]; however approaches exist for circumventing these shortcomings [11].

On the other hand, the advantages of the SVM approach include [74]:

1. The training phase is a convex quadratic programming problem. This is an advantage if we consider that there exist computationally efficient algorithms that can be applied for this matter, with which the finding of the global extremum is guaranteed.

2. This algorithm does not face the problem of overfitting, as neural networks do.

3. It allows to working with nonlinear relationships between the data, since it generates nonlinear functions by means of kernels.

4. It generalizes well, even with a small number of training samples.

5.4 Discriminant Analysis

In the generic classification problem, the outcome of interest G falls into J unordered classes, and the objective is to build a rule for predicting the class membership of an item based on r measurements of predictors or features $\mathbf{X} \in R^r$; the training sample consists of the class membership and the predictors for N items. To distinguish the known classes from each other, we associate a unique *class label* (or output value) with each class; the observations are then described as *labeled observations*.

Linear discriminant analysis (LDA) is a well-known method for classification, in which a test observation with predictor $\mathbf{X_0}$ is classified to the class with centroid closest to $\mathbf{X_0}$, where distance is measured in the *Mahalanobis* metric using the pooled within-group covariance matrix. This procedure can be justified by assuming that the predictors have a multivariate Gaussian distribution, with different means but a common covariance matrix among the classes. The observation is then assigned to the class having the maximum posterior class probability; this results in the rule described earlier if the class prior probabilities are the same. Characteristics of this procedure include [35]:

1. All the relevant distance information is contained in the at most $J-1$-dimensional subspace of R^r spanned by the J group centroids.

2. Decision boundaries are linear.

3. In reduced versions of LDA due to Fisher and Rao, in which dimensions are reduced, the data can be plotted in the reduced space, giving a graphical representation of the group separation. Often two or three dimensions are needed, even for large J.

4. This dimension-reduced model can show better classification performance, even when having many classes and limited training data, since the reduced space can be more stable and yield improved misclassification results on test data.

Let us consider the binary classification problem, addressed with LDA [48, 34], where we wish to discriminate between two classes Π_1 and Π_2. Let

$$\mathrm{P}(\mathbf{X} \in \Pi_i) = \pi_i, i = 1, 2, \tag{5.5}$$

be the *prior probabilities* that a randomly selected observation $\mathbf{X} = \mathbf{x}$ belongs to either Π_1 or Π_2. Suppose also that the conditional multivariate probability density of $\mathbf{X}$ for the ith class is

$$\mathrm{P}(\mathbf{X} = \mathbf{x}|\mathbf{X} \in \Pi_i) = f_i(\mathbf{X}), i = 1, 2. \tag{5.6}$$

Note that there is no requirement that the $f_i(\cdot)$ be continuous; the could be discrete or be finite mixture distributions or even have singular covariance matrices. From equations (5.5) and (5.6), Bayes' theorem yields the *posterior probability*:

$$p(\Pi_i|\mathbf{x}) = \mathrm{P}(\mathbf{X} \in \Pi_i|\mathbf{X} = \mathbf{x}) = \frac{f_i(\mathbf{x})\pi_i}{f_1(\mathbf{x})\pi_1 + f_2(\mathbf{x})\pi_2}, \tag{5.7}$$

that the observed $\mathbf{x}$ belongs to Π_i, $i = 1, 2$.

For a given $\mathbf{x}$, a reasonable classification strategy is to assign $\mathbf{x}$ to that class with the higher posterior probability. This strategy is called the **Bayes' rule classifier**. Formally, we assign $\mathbf{x}$ to Π_1 if

$$\frac{p(\Pi_1|\mathbf{x})}{p(\Pi_2|\mathbf{x})} > 1, \tag{5.8}$$

and we assign $\mathbf{x}$ to Π_2 otherwise. The ratio $p(\Pi_1|\mathbf{x})/p(\Pi_2|\mathbf{x})$ is referred to as the *odds-ratio* that Π_1 rather than Π_2 is the correct class given the information in $\mathbf{x}$. Substituting equation (5.7) into (5.8), the Bayes' rule classifier assigns $\mathbf{x}$ to Π_1 if

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > \frac{\pi_2}{\pi_1}, \tag{5.9}$$

and to Π_2 otherwise. On the boundary $\{\mathbf{x} \in R^r | f_1(\mathbf{x})/f_2(\mathbf{x}) = \pi_2/\pi_1\}$, we randomize between assigning $\mathbf{x}$ to either Π_1 or Π_2.

Now, we make the Bayes' rule classifier more specific by following Fisher's assumption that both multivariate probability densities in equation (5.6) are multivariate Gaussian having arbitrary mean vectors and a common covariance matrix. That is, we take $f_1(\cdot)$ to be a $\mathbf{N}_r(\mu_1, \mathbf{\Sigma}_1)$ density and $f_2(\cdot)$ to be a $\mathbf{N}_r(\mu_2, \mathbf{\Sigma}_2)$ density, and we make the homogeneity assumption that $\mathbf{\Sigma}_1 = \mathbf{\Sigma_2} = \Sigma_{XX}$.

The ratio of the two densities is given by

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = \frac{\exp\{-\frac{1}{2}(\mathbf{x} - \mu_\mathbf{1})^\tau \Sigma_{XX}^{-1}(\mathbf{x} - \mu_\mathbf{1})}{\exp\{-\frac{1}{2}(\mathbf{x} - \mu_\mathbf{2})^\tau \Sigma_{XX}^{-1}(\mathbf{x} - \mu_\mathbf{2})}, \tag{5.10}$$

where the normalization factors $(2\pi)^{-r/2}\,|\mathbf{\Sigma_{XX}}|^{-1/2}$ in both numerator and denominator cancel due to the equal covariance matrices of both classes. Taking logarithms (a monotonically increasing function) of equation (5.10), we have that

$$\log_e \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = (\mu_1 - \mu_2)^\tau \mathbf{\Sigma}_{XX}^{-1}\mathbf{x} - \frac{1}{2}(\mu_1 - \mu_2)^\tau \mathbf{\Sigma}_{XX}^{-1}(\mu_1 + \mu_2) \tag{5.11}$$

The second term in the right-hand side of equation (5.11) can be written as

$$(\mu_1 - \mu_2)^\tau \mathbf{\Sigma}_{XX}^{-1}(\mu_1 + \mu_2) = \mu_1^\tau \mathbf{\Sigma}_{XX}^{-1}\mu_1 - \mu_2^\tau \mathbf{\Sigma}_{XX}^{-1}\mu_2. \tag{5.12}$$

It follows that

$$L(\mathbf{x}) = \log_e \left\{ \frac{f_1(\mathbf{x})\pi_1}{f_2(\mathbf{x})\pi_2} \right\} = b_0 + \mathbf{b}^\tau \mathbf{x} \tag{5.13}$$

is a linear function of $\mathbf{x}$, where

$$\mathbf{b} = \mathbf{\Sigma_{XX}^{-1}}(\mu_1 - \mu_2) \tag{5.14}$$

$$b_0 = -\frac{1}{2}\left\{\mu_1^\tau \mathbf{\Sigma}_{XX}^{-1}\mu_1 - \mu_2^\tau \mathbf{\Sigma}_{XX}^{-1}\mu_2\right\} + \log_e(\pi_2/\pi_1). \tag{5.15}$$

Thus, we assign $\mathbf{x}$ to Π_1 if the logarithm of the ratio of the two posterior probabilities is greater than zero; that is,

$$\text{if } L(\mathbf{x}) > 0, \text{assign } \mathbf{x} \text{ to } \Pi_1. \tag{5.16}$$

Otherwise, we assign $\mathbf{x}$ to Π_2. Note that on the boundary $\{\mathbf{x} \in R^r | L(\mathbf{x}) = 0\}$, the resulting equation is linear in $\mathbf{x}$ and, therefore, defines a hyperplane that divides the two classes. The rule in equation (5.16) is generally referred to as ***Gaussian Linear Discriminant Analysis (LDA)***.

The part of the function $L(\mathbf{x})$ in equation (5.13) that depends upon $\mathbf{x}$,

$$U = \mathbf{b}^\tau \mathbf{x} = (\mu_\mathbf{1} - \mu_\mathbf{2})^\tau \mathbf{\Sigma_{XX}^{-1}}\mathbf{x}, \tag{5.17}$$

is known as *Fisher's linear discriminant analysis (LDF).*

Finally, many techniques are based on models for the class densities [34]:

- Linear and quadratic discriminant analysis use Gaussian densities;
- More flexible mixtures of Gaussians allow for nonlinear decision boundaries;
- General nonparametric density estimates, for each class density, allow the most flexibility;
- *Naive Bayes* models are variant of the previous case, and assume that each of the class densities are products of marginal densities; that is, they assume that the inputs are conditionally independent in each class.

The analysis of the suitability of the previous techniques within a given problem, is justified because, in practice, linear decision boundaries are often too crude and nonlinear boundaries can be more effective. This research work will start designing a linear discriminant analysis algorithm, as a solution for the problem being addressed; afterwards, other discriminant analysis methods will be analyzed, in order to try to enhance the classification task. Formally, those methods are:

- **Flexible Discriminant Analysis (FDA).** Formulates LDA as a linear regression problem and then, abandons the idea of the regression task to substitute it by a non parametric one, thus, enlarging the basis of the feature space.
- **Penalized Discriminant Analysis (PDA).** Considers an explicit expansion of the feature space and penalizes rough coordinates. It is particularly useful when the problem involves a large amount of predictors or features.
- **Mixture Discriminant Analysis (MDA).** Extends the LDA by allowing to have multiple class prototypes (the LDA considers a single class prototype) by representing each class by a gaussian mixture.

These algorithms represent promising alternatives that can be considered in future stages of this research work, in order to circumvent issues related to the LDA. However, the decision of selecting one of them as part of the experimentation platform will depend on a further analysis of the problem and its characteristics, which will be conducted later on the research time line; in this way, one or more of them (or even other DA approaches) can be taken into account, in order to select the one(s) which best fit according to the problem.

5.5 k-Nearest Neighbors

The k-Nearest Neighbors (k-NN) [35] is a non-parametric method used for both classification and regression problems which does not require to fit a model to the input data. In k-NN classification, this algorithm uses a query vector to find the k data points in the training set that are closest in *diantance* to the query and finally classifies the new data point by a majority vote among the k neighbors or, in other words, the class most frequently represented among the neighbours. Ties, if any, are broken randomly and k is typically a small positive integer; in the case where $k = 1$ the new instance or object is simply assigned to the class of that single nearest neighbor. On the other hand, the output of k-NN regression is a property value for the new instance, based on the average of the values of its neighbors.

There are several distance metrics that can be considered in a k-NN classifier, including the euclidean distance, cosine similarity, manhattan distance, among others. The use of any of these will define a specific *similarity space* and the neighborhood of the query point may change between different metrics.

It is a simple algorithm, but has proved to be effective in different classification and regression problems, like recognizing handwritten digits, classification of digital images and forecasting scenarios with several applications. It is also successful in problems where classes have several possible prototypes and the boundary between classes is irregular [9].

The best choice of k depends upon the data. Generally, larger values reduce the effect of noise on the classification, but make boundaries between classes less distinct. A good k can be selected by various heuristic and optimization techniques that include evolutionary exploration, grid search (an exhaustive search of a manually specified subset of values), random search and bootstrap methods. In binary (two class) classification problems, it is helpful to set k to be an odd number as this avoids tied votes.

Furthermore, the accuracy of k-NN can be degraded by the presence of noisy or irrelevant features, or if the feature weights are not consistent with their importance. Several research efforts have been conducted on exploring the feature space in the search for an optimal subset that can be fed to the classifier and finally improve its performance; in this case, evolutionary algorithms (Genetic Algorithms) and statistical methods (Principal Component Analysis, for instance) are popular approaches.

Figure 5.2 depicts an example of how k-NN assigns new instances to a given class. It is a binary classification problem in which one class is represented by squares and the other by triangles; the green circle represents a test instance which should be classified as being either from one class or the other. We can see that if we set $k = 3$ (inner circle) the new instance is assigned to the class represented by triangles, since there are 2 of them and only 1 square inside that circle. On the other hand, if we set $k = 5$ (outer circle) it is assigned to the squares class, because the number of elements of this class exceeds that of the one represented by triangles.

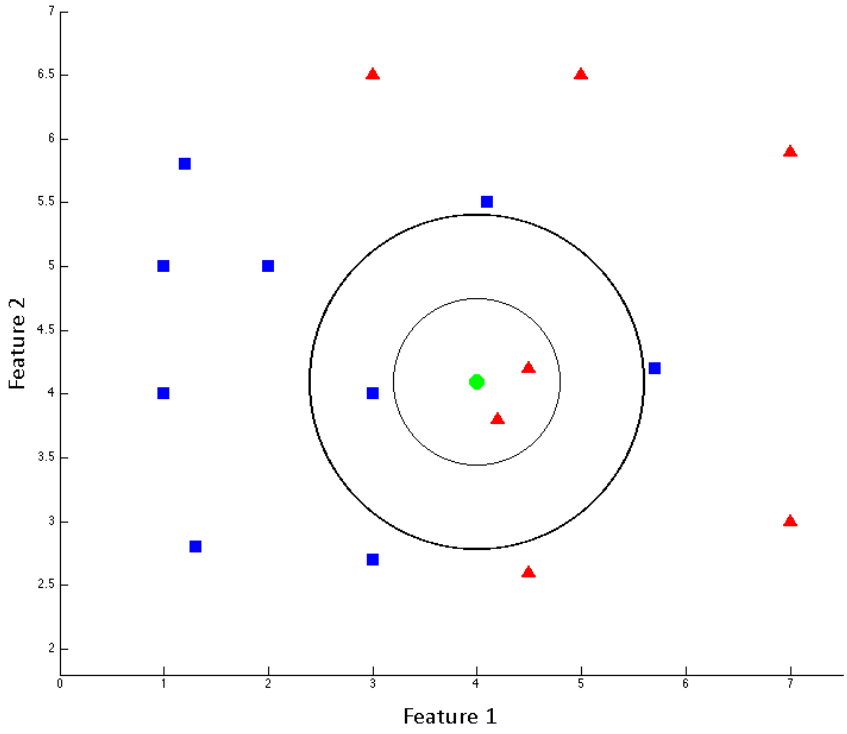


Figure 5.2: k-NN Classification Example

5.6 Case-based Reasoning

Case-based Reasoning (CBR) is an Artificial Intelligence (AI) technique, which has the objective to support the capability of reasoning and learning in advanced decision support systems [57]. It is a reasoning paradigm that exploits the specific knowledge collected on previously solved situations, known as *cases*.

The basic idea is to solve new problems by identifying, retrieving and reusing previous similar cases, based on the heuristic principle that similar problems have a high likelihood of having similar solutions. Within an approach like this, problems are solved by adapting the solutions that were previously used to solve similar historical problems.

The CBR cycle is composed of four basic tasks: retrieve, reuse, revise and retain. Typically, when a new case is provided to query the system, it is pre-processed in order to construct an appropriate representation of it, which is then interpreted in the reasoning cycle, considering that the abstraction of the original problem should match the knowledge representation of the case base. The inner workings of the CBR stages are described as [2]:

1. **Retrieve:** A similarity-search is conducted over the database for the set of cases which are most similar to the query case. In this stage the system applies a similarity measure between the new case and those stored in the database. One of the most common similarity-search methods is the *nearest neighbors*, in which typically the similarity between the query case and those stored in the database is determined by a metric such as cosine similarity, manhattan distance, correlation factor, among others, which are calculated upon the vector of visual features that were extracted from the images. In medical-imaging applications, the retrieval of similar cases is performed by implementing Content-Based Image Retrieval (CBIR) on a database that contains, both, images and the metadata that are used to compute similarity-search [72].

2. **Reuse:** Once the set of similar cases has been retrieved, the system reuses the information of one or more of them, by way of a interpretation or possible adaptation, in order to provide a solution to the new problem which is the main objective of this stage; the approach that will be used to reuse the information of retrieved cases depends on the nature of the problem that is going to be solved, being adaptation the most frequently used in diagnosis tasks. However, not all CBR systems provide a suggested solution, but focus only in retrieving the most relevant cases, as a form case-based retrieval systems.

3. **Revise:** This step evaluates the solution suggested by the system in a real world scenario or against the revision of an expert which will reject, correct or confirm it. In CBR both correct and incorrect solutions are equally important, since they represent the experience that is drawn from by the system from the its knowledge base, but the latter have to be identified and repaired to prevent failing again in future.

4. **Retain:** The revised solution is stored as a new case in the database for future problem solving. This is why it is important to validate the fitness of the proposed solutions; otherwise, reusing historical information would not be useful to compute a solution and the accuracy of the system would be compromised. With this process the ground truth of the system increases which results in the fact that subsequent queries are solved over a broader experience.

The previous steps are depicted in Figure 5.3, where it can be observed that, at the beginning of the process, a new *problem* defines a new *case*, which is then used to *retrieve* one or more cases from the collection of *previous cases*. Afterwards, the retrieved cases are *reused* and combined with the new case into a *solved case* (i.e. a proposed solution to the problem). The fitness of the solution is tested

in the *revise* process, by applying it in a real world scenario or evaluated by a *teacher*, and repaired if failed. The experience drawn from this solution is *retained* in the case base, in order to be reused in the future, by storing it as a *new case* or by modifying existing cases.

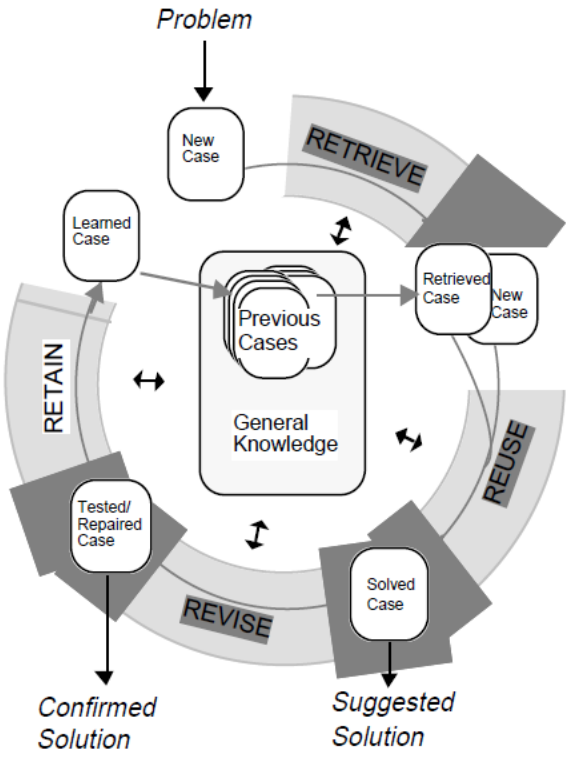


Figure 5.3: The Case-Based Reasoning Cycle [2].

Content-Based Image Retrieval systems have been applied in different clinical applications [1], including breast cancer diagnosis by mammography [104], in which the fundamental process is to retrieve historical images containing lesions that are similar to the ones detected in a query image provided by a physician and, also, reuse them for classifying lesions [51, 100].

The earliest CBR system reported in literature is CYRUS [55, 56], developed in 1977, which had an indexed memory structure and employed knowledge as cases. Afterwards, other systems such as CASEY [58] and MEDIATOR [86] were implemented based on CYRUS.

CBR is different of other AI methods, like Neural Networks or Rule-based Systems, in that it does not aim to generalize from the examples shown. Rather, this method keeps and exploits the specific instances of problems that have been collected in the past, and which have a known solution. In classical CBR no efforts are made to extract more abstract information; i.e. there is no attempt to elicit rules or models from cases. This way, the difficulties of knowledge acquisition and of knowledge representation are decreased, since specific cases are re-used directly as reference for decision making under a query problem.

As the number of cases grows, more representative examples can be retrieved, making it easier to find a suitable solution to the present problem. However, it should be noticed that proper case base maintenance policies must be defined, in order to establish when to delete, retain and/or update its elements.

We can see at this point that CBR is not only a computer reasoning method, but also a common human problem-solving behavior we can come across in everyday life. It is inspired by the cognitive

model through which humans solve problems, based on previous experience adapted to the current situation. For instance, the medical domain is quite suitable for a CBR system, since medical doctors often recall similar cases that they have learned and adapt them to the situation they are analyzing at a given moment.

Moreover, according to [4], the motivations to apply CBR method in the medical domain are:

1. The CBR method can solve a problem in a way similar to the normal behavior of human problem solving e.g. it solves a problem using experience.
2. A CBR system could be of great value for a less experienced person because the case library can be used as knowledge.
3. A CBR system can start working with few reference cases in its case base and then learn though the time by adding new cases into the repository. This is an analogy of doctors or engineers starting their practice with a few cases and then gradually increasing their experiences.
4. A CBR system can provide more than one alternative for a similar problem which is beneficial for the clinician.
5. CBR can help to reduce the recurrence of a wrong decision because the case library could contain both success and failure of cases.
6. Knowledge elicitation is most of the time a bottleneck in the health science domain since human behavior is not always predictable. The CBR method can overcome this because prediction is based on the experience or old cases.
7. It is useful if the domain is not clear enough, since CBR does not depend on any rules or any models.
8. Systems using CBR can learn new knowledge by adding new solved cases into the case library, so domain knowledge is also updating in time.

On the other hand, there are some major challenges that researchers have to consider when designing and implementing a medical application of a CBR model, according to Ahmed et al [4]:

1. **Limited number of reference cases.** Even though a CBR system can work with a few number of reference cases, the performance might not be the best.
2. **Feature extraction.** Since in CBR cases are described by a feature vector, this may become an issue when this process has to be performed on complex data formats, such as images, sensor signals, etc.
3. **Adaptation.** The medical domain is often subject to complex knowledge and recommendations change through the time, and, since cases often consists of large number of features, it is a major challenge to apply knowledge discovery and automatic adaptation strategies in this area.

As the reader may expect, the CBR methodology is going to be followed in this research work, since one of the main objectives is to build a CAD system which considers similar historical cases, with known pathologies, to provide an answer to the end-user, who finally revises this output and determines the correct answer. Finally, this answer is stored in the database of mammographic lesions for future usage, as a process to retain this new cases.

5.7 Fuzzy Unordered Rules Induction Algorithm (FURIA)

The Fuzzy Unordered Rules Induction Algorithm (FURIA) [46, 47] is an extension of RIPPER [23] which is a the state-of-the-art rule learner. This section provides a brief summary of this algorithm's internal processes; readers are encouraged to refer to the cited articles for a complete description of this classifier.

The novelty of FURIA is that instead conventional rules it learns fuzzy rules and unordered rule sets instead of rule lists. The importance of using fuzzy rules is that they avoid *sharp* decision boundaries and, consequently, softens the transitions between different classes. With fuzzy rules the support for a given class decreases gradually from completely belonging to it to not being related to that class at all.

Conventional rule learners build a *hierarchical* decision list. In this procedure, these algorithms generate rules for every class starting with the one that is less frequent in the dataset and ending with the second more frequent one. Also, they add a default rule for the majority class and query cases are then classified using the first rule that covers the new instance. This mechanism possess some disadvantages since it can get biased towards the default class because classes are not treated in a symmetric way. To avoid this problem, FURIA produces an unordered set of rules for each class in a one-vs-the-rest fashion.

However, because of this process the model is not necessarily complete and, consequently, some queries would not be covered by any of the generated rules. Therefore, this algorithm also conducts a rule stretching method in which existing rules are generalized until they cover the new example. This represents an advantage against using default rules in the sense it is a local strategy that exploits information in the vicinity of the query instance. Generalizing or *stretching* a rule is conducted by deleting all the antecedents until it covers the new instance and a generalization is considered minimal if it does not delete more antecedents than necessary to cover the query. After all minimal generalizations are computed, this algorithm re-evaluates each rule with the Laplace accuracy on the training data and, finally, classifies the query with the rule that possesses the highest evaluation.

Another modification that FURIA performs tot the RIPPER algorithm is that it does not perform any pruning since the authors found out that it negatively affected the performance of FURIA. Instead they used the complete training data to learn the initial rule set; by doing this, more specific rules are produced by the underlying rule generator in RIPPER. Additionally, small rules represent a good starting point for fuzzification, which is a process that makes rules more general but not more specific. However, FURIA still performs pruning during the optimization stage, where a replacement and a revision rule is created, unless the pruning mechanism tries to remove all antecedents of a given rule in which case the original rule is used for the comparison in the optimization phase.

Fuzzy rules in FURIA are obtained by replacing antecedent intervals by fuzzy sets with trapezoidal membership functions, using a greedy algorithm that extends the support of a given rule, based on the improvement of a *purity* factor that measures the component-wise confidence of the fuzzy rule. For the fuzzification of a given antecedent $(A_i \in I_i)$, where I_i is an interval of the original rules obtained by RIPPER which are taken as cores of the pursued fuzzy intervals I_i^F, it ignores those instances excluded by any other antecedent $(A_j \in I_j^F)$, with $j \neq i$, to keep only the relevant training data D_{T}^i from the training set D_{T}; formally:

$$D_{\mathrm{T}}^i = \left\{ \mathbf{x} = (x_1 ... x_k \in D_{\mathrm{T}}) | I_j^F(x_j) > 0 \text{ for all } j \neq i \right\} \subseteq D_{\mathrm{T}} \tag{5.18}$$

where $\mathbf{x}$ is a k-dimensional attribute vector in the training set.

Then, it divides the training data D_{T}^i in the subset of positives instances $D_{\mathrm{T}^+}^i$ and the subset of negative instances $D_{\mathrm{T}^-}^i$ and measures the quality of the resulting fuzzification by computing the purity of the rule as follows:

$$\text{pur} = \frac{p_i}{p_i + n_i} \tag{5.19}$$

where

$$p_i = \sum_{\mathbf{x} \in D^i_{\mathrm{T}+}} \mu_{A_i}(\mathbf{x}) \tag{5.20}$$

$$n_i = \sum_{\mathbf{x} \in D^i_{\mathrm{T}-}} \mu_{A_i}(\mathbf{x}) \tag{5.21}$$

In the previous equations, p_i can be regarded as the summation of the membership values of each attribute vector $\mathbf{x}$ in the subset of positive instances of the training set, relative to a given antecedent A_i. Naturally, the n_i is the summation of membership values of the negative instances in the training set.

When the algorithm has learned the fuzzy rules $r_1^{(j)} \dots r_k^{(j)}$ for a given class λ_j, the support of the class for a given query instance $\mathbf{x}$ is computed as follows:

$$s_j(\mathrm{x}) = \sum_{i=1 \dots k} \mu_{r_i^{(j)}}(\mathrm{x}) \cdot CF\left(r_i^{(j)}\right) \tag{5.22}$$

where $CF\left(r_i^{(j)}\right)$ is a certainty factor of rule $r_i^{(j)}$ which is calculated as follows:

$$CF\left(r_i^{(j)}\right) = \frac{2\frac{\left|D_{\mathrm{T}}^{(j)}\right|}{|D_{\mathrm{T}}|} + \sum_{\mathbf{x} \in D_{\mathrm{T}}^{(j)}} \mu_{r_i^{(j)}}(\mathbf{x})}{2 + \sum_{\mathbf{x} \in D_{\mathrm{T}}} \mu_{r_i^{(j)}}(\mathbf{x})} \tag{5.23}$$

where $D_{\mathrm{T}}^{(j)}$ is the subset of instances in the training set labeled with class λ_j.

Finally, the class with the maximal support is assigned to the query instance and, in case none of the rules is able to cover it, i.e. $s_j(\mathrm{x}) = 0$ for all classes, FURIA performs the rule stretching mechanism.

5.8 Classifier Ensembles

A classifier ensemble is a group of classifiers that are combined with the objective of improving the performance of a single classifier. When a new instance has to be classified, it is fed to each of the members of the ensemble and then their outputs are appropriately aggregated to assign a class to the new instance. The result is a new classifier that outperforms all of its individual members.

It is important for each of the considered classifiers to be different from one another, because the ensemble will take advantage of the diversity of its members to improve their overall generalization ability. If they have been trained with different datasets, different features or represent a variant of the same classifier, then the misclassified instances will not be necessarily the same and, at the end of the process, those errors can be reduced by combining the answers of the rest of the classifiers. In this section we describe two alternatives to build classifier ensembles that we later used in one stage of our experiments: Bagging and Random Subspace.

5.8.1 Bagging

The concept of *Bagging* was first introduced by Breiman as an acronym for *Boostrap AGGregatING* [12], in which a series of classifiers are built on bootstrap replicates of the training set and then used as new learning sets. Outputs are computed by a plurality vote if a label has to be predicted, or by computing the average over the obtained versions when predicting a numerical output. This method is based on the premise that if perturbing the training set can cause significant changes in the classifier, then bagging can improve its accuracy.

Using different training sets generated at random from the same distribution of the problem and building different classifiers with these training sets should introduce diversity to the ensemble. However, in reality we do not have access to a sequence of replicates of the learning set. On the contrary, we only get one labelled dataset and, consequently, have to imitate the process of randomly generating replicates of it. In Bagging, this process is done by sampling with replacement from the original learning set to obtain the new instances.

Moreover, this method takes advantage of a unstable classifiers, i.e. those which produce very different outputs with small changes in the training set. Base classifiers used under Bagging should have this property to preserve the diversity of the ensemble because, otherwise, the process would end up building very similar, or identical, classifiers that would not provide the possibility to improve the performance of a single classifier.

Algorithm 1 depicts the Bagging method. It is composed of two stages: training and classification. Bagging is a parallel algorithm in both training and classification stages and, thus, its members can be trained at the same time. The ensemble is generated in the training phase by building a given number of classifiers with a bootstrap sample taken from the original training set. On the other hand, in the classification process the algorithm uses each of the members of the ensemble to classify a given input and, finally, counts the number of votes that each class received from the classifiers. The class that was voted the most is assigned to the input instance.

Algorithm 1 Bagging algorithm

Require: L, the number of classifiers to train; $\mathbf{T}$, the training set.

1: **procedure** TRAINING(L, $\mathbf{T}$) ▷ **Training phase**
2: $E \leftarrow \varnothing$ ▷ Initialize E, the ensemble
3: **for** $k = 1, \ldots, L$ **do**
4: Take a bootstrap sample S_k from $\mathbf{T}$
5: Build a classifier E_k using S_k as the training set
6: $E = E \cup E_k$ ▷ Add the classifier to the ensemble
7: **end for**
8: **return** E
9: **end procedure**

Require: E, the ensemble; $\mathbf{x}$, the instance to be classified.

10: **procedure** CLASSIFICATION(E, $\mathbf{x}$) ▷ **Classification phase**
11: Apply $E_1, \ldots, E_L$ on $\mathbf{x}$
12: **return** The class with the maximum number of votes
13: **end procedure**

5.8.2 Random Subspace

Classifiers in an ensemble can also be built with different subsets of features. In Random Subspace, first introduced by Ho [44], a subset of features of a predefined size d is randomly chosen and then used to build the classifiers. The author claims that, considering decision trees, good results are obtained with $d \approx n/2$, being n the total number of features. This method is useful when redundant information is *dispersed* across the set of features as opposite to being concentrated in a subset and it has proved to maintain high accuracy on training data and an improved generalization performance in terms of accuracy.

The validity of this algorithm does not depend on the construction method of the underlying base classifier. It can be used for different types of tree-based algorithms, supervised or unsupervised clustering and even SVMs. Actually, it is a generalization of the Random Forest algorithm, since it is not necessarily composed of decision trees but of any type of classifier.

For a given n-dimensional feature space, there are 2^n possible subsets that can be selected and with each of them a classifier can be built. The random selection of these features enables the algorithm to conveniently explore many possibilities and if different feature dimensions are considered for building the classifiers, the diversity of the ensemble can also be increased.

Moreover, while other classification schemes are negatively affected by the *curse of dimensionality*, Random Subspace takes advantage from it, since high dimensional feature spaces provide a vast number of subspaces that can be selected to train a classifier, each one of which will generalize their classification in a different way, provided that they are trained with a different subset of features, resulting in an classifier ensemble that is likely to achieve a high performance, if there are no ambiguities in the feature set.

Random Subspace is also a parallel algorithm and, therefore, the generation of each of the members of the ensemble can be done at the same time in different processors, allowing to use this algorithm for real-world implementations that require a fast-learning method that can be built and operated rapidly. The outputs of each classifier in the ensemble are combined to compute the final answer of the algorithm, typically with a majority vote mechanism. Algorithm 2 depicts the internal working of this method.

5.9 Learning from imbalanced datasets

A dataset is referred to as being *imbalanced* when it contains an unequal distribution between its classes, i.e. it has many more instances from certain classes than others [50, 19, 20]. This poses a problem to several machine learning algorithms (such as neural networks, SVMs, decision trees, among others) that assume a relatively balanced distribution between classes and, therefore, typically fail in accurately learning the *underrepresented* class. For some applications, this is a serious problem, since the correct classification of the minority class is often very critical as it is the case of medical diagnosis where positive or disease cases are rare in a normal population and identifying them with high precision can save the patient's life. We can measure how skewed a given dataset is, by computing its *imbalance ratio*, taken as:

$$\text{IR} = \frac{\#\text{ positive instances}}{\#\text{ negative instances}} \tag{5.24}$$

There are three important issues of this problem that researchers have studied: (1) the nature of the imbalance issue, in order to discover which domains affect *standard* classifiers the most, (2) the possible solutions that could be implemented to solve the class-imbalance problem and (3) the appropriate performance metrics to consider in this problem, since measuring the accuracy of a classifier is

Algorithm 2 Random Subspace algorithm

Require: L, the number of classifiers to train; $\mathbf{T}$, the training set; t, the number of features to consider.
Ensure: $t < |\mathbf{T}|$

procedure TRAINING(L, $\mathbf{T}$, t) ▷ **Training phase**
  $E \leftarrow \varnothing$ ▷ Initialize E, the ensemble
  for $k = 1, \ldots, L$ **do**
    Create a training set S_k by randomly chosing t features from $\mathbf{T}$
    Build a classifier E_k using S_k as the training set
    $E = E \cup E_k$ ▷ Add the classifier to the ensemble
  end for
  return E
end procedure

Require: E, the ensemble; $\mathbf{x}$, the instance to be classified.
procedure CLASSIFICATION(E, $\mathbf{x}$) ▷ **Classification phase**
  Apply $E_1, \ldots, E_L$ on $\mathbf{x}$
  return The class with the maximum number of votes
end procedure

no longer a good approach because the minority class has a low impact on the accuracy as compared to the majority class.

A skewed data distribution between classes is not the only problem that may negatively affect the performance of a classifier. Other issues that may impact the classification precision, which are also related to the imbalanced problem, include [89, 37]:

1. ***Small sample size.*** If the size of the training dataset increases, the large error rates induced by the class-imbalanced issue decreases. This means that the sole problem an imbalanced class distribution will not necessarily affect a classifier's performance if enough data is provided.

2. ***Class separability.*** If the boundaries between any two classes are overlapped, the imbalance problem will significantly affect the number of correctly-classified examples of the minority class. On the other hand, separable domains will not be sensitive to any imbalance ratio.

3. ***Within-class concepts.*** When a single class is decomposed in various subclasses, or subconcepts, typically instances from different subclasses are sampled into the training dataset and very often they do not contain the same number of examples from one subclass to the other. This problem is know as *within-class imbalance* and refers to the skewed distribution of subclasses within a given class, resulting in an increased learning complexity that affects the performance.

Moreover, three main research avenues have been explored for solving the problem of imbalanced class distribution:

1. ***Algorithm-level*** approaches that try to modify or adapt the internal processes of existing classifier algorithms to enable them to deal with class-imbalance datasets by biasing the learning process towards the minority class. A thorough

2. ***Data-level*** methods that resample the training data to eliminate the skewed class distribution. They are independent from the underlying classifier because they perform a data pre-processing step that avoids the modification of the selected algorithm and are, consequently, more versatile, i.e. they can be used, equally, with different classification methods. These techniques either *undersample* the majority class or *oversample* the minority or underrepresented one.

3. ***Cost-sensitive learning*** techniques that introduce misclassification costs to instances at data-level but also modifications to the learning process of the classification algorithm to enable it to take a cost-matrix that determines the penalties that it will receive for incorrectly classifying instances. Usually, this cost-matrix will bias the learning process towards the minority class by drastically penalizing the classifier if it fails to correctly classify instances from that class and, very often, considering also a soft penalization for *correctly* classifying instances from the *majority* class, with the objective of preventing this class from taking over the learning process.

Preprocessing, or data-level, techniques can be easily integrated with any classifier algorithm. Those techniques can be divided into three categories: (1) *undersampling* methods that eliminate instances from the original training set, usually belonging to the majority class, (2) *oversampling* methods which build a new training set by creating new instances based on the existing ones that are related to the minority class and (3) *hybrid* techniques that combine the previous two methods.

In this research work we used a combination of two preprocessing methods in our experiments: *Synthetic Minority Oversampling technique* (SMOTE) [18] and Wilson's ENN rule [101]. In this hybrid approach, referred to as SMOTE+ENN, the SMOTE algorithm oversamples the training set to introduce new synthetic, or artificial, instances from the minority class by interpolating several examples from that class that are *close* to each other in feature space and the ENN rule is a simple method which undersamples the resulting dataset by removing any instance whose class differs from that of its three nearest neighbors.

Specifically, for the subset of instances belonging the minority class $S_{minority} \in S$, where S is the original training set, the algorithm computes the k-Nearest Neighbors for each instance $\mathbf{x_i} \in S_{minority}$, for a specified integer k. Then, to create a new instance the algorithm randomly selects one of those k instances, computes the difference between such vector and $\mathbf{x_i}$, multiplies the resulting difference by a random number between $[0, 1]$ and finally adds this vector to $\mathbf{x_i}$. Formally:

$$\mathbf{x_{new}} = \mathbf{x_i} + (\hat{\mathbf{x}}_\mathbf{i} - \mathbf{x_i}) \cdot \delta \tag{5.25}$$

where $\mathbf{x_i} \in S_{minority}$ is the instance from the minority class that will be replicated, $\hat{\mathbf{x}}_\mathbf{i}$ is one of the k-Nearest Neighbors and $\delta \in [0, 1]$.

Afterwards, the ENN rule goes though all the instances $\mathbf{x_j} \in S_{SMOTE}$, where S_{SMOTE} is the resulting set after applying SMOTE, and eliminates $\mathbf{x_j}$ if its class is different from the class of its three nearest-neighbors in S_{SMOTE}.

5.10 Measuring the Performance of Classification Algorithms

Evaluation metrics play an important role in assessing classification performance, but also in guiding the classifier modeling. As the number of classification methods increases in the research community, having standardized evaluation metrics is critical not only to appropriately measure the performance of the algorithm, but also to objectively compare results obtained with different methods and being able to draw conclusions.

In binary classification problems, the *confusion matrix*, or contingency table, allows the visualization of a classifier's performance. As it is shown in Table 5.1 each column of the matrix represents

the predicted classes while each row represents the actual class of all instances that were processed by the classifier. Four important metrics are obtained: the number of true positives (TP), false negatives (FN), false positives (FP) and true negatives (TN).

	Predicted as positive	Predicted as Negative
Actually Positive	True Positives (TP)	False Negatives (FN)
Actually Negative	False Positives (FP)	True Negatives (TN)

Table 5.1: The confusion matrix.

Moreover, several performance metrics can be derived from the confusion matrix:

$$accuracy = \frac{TP + TN}{TN + FP + FN + TP} \tag{5.26}$$

$$\text{True Positive Rate: } TP_{rate} = \frac{TP}{TP + FN} \tag{5.27}$$

$$\text{True Negative Rate: } TN_{rate} = \frac{TN}{TN + FP} \tag{5.28}$$

$$\text{False Positive Rate: } FP_{rate} = \frac{FP}{TN + FP} \tag{5.29}$$

$$\text{False Negative Rate: } FN_{rate} = \frac{FN}{TP + FN} \tag{5.30}$$

$$\text{Positive Predictive Value: } PP_{value} = \frac{TP}{TP + FP} \tag{5.31}$$

$$\text{Negative Predictive Value: } NP_{value} = \frac{TN}{TN + FN} \tag{5.32}$$

The true positive rate, stated in equation 5.27, and the true negative rate, in equation 5.28, are more commonly referred to as sensitivity and specificity, respectively. Moreover, the accuracy of the classifier, computed by equation 5.26, is the most frequently used performance metric. However, in a class-imbalanced scenario, accuracy is no longer a proper measure since the minority class has a low impact on that metric as compared to that of the prevalent class. For example, if the minority class represents only 1% of the data, a classifier that assigns all new instances to the majority class will achieve a 99% accuracy; this means that it fails to classify the minority class, which is of major importance for several applications, even though its accuracy is very high.

Therefore, more appropriate metrics have been designed for class-imbalanced problems and are frequently adopted by the research community to provide a comprehensive assessment of performance in this scenario. The most common are the geometric mean ($G-mean$) and the area under the Receiver Operating Characteristic (ROC) curve (AUC):

$$G - mean = \sqrt{TP_{rate} \cdot TN_{rate}} \tag{5.33}$$

$$\text{AUC} = \frac{1 + TP_{rate} - FP_{rate}}{2} \tag{5.34}$$

$G-mean$ measure the balanced performance of learning algorithms when both the true positive rate and the true negative rate are expected to be high simultaneously. It provides a result that stays closer to the smaller numbers that were considered for computing this metric, as opposed to the arithmetic mean which stays always in the middle. For example, the arithmetic mean of 1 and 5 would be 3, while their $G-mean$ would be 2.236, which is closer to the smaller number. Therefore, this metric will be high when both the sensitivity and specificity of the classifier are high or when the difference between them is small.

On the other hand, AUC describes a trade-off between the true positive and the false positive rates, which can be regarded as a trade-off between benefits and costs. In the ROC curve, sensitivity is plotted on the y-axis and 1-Specificity is plotted along the x-axis. The larger the area (total area is 1.0) the better the classification is. For instance, in the case of AUC$= 1.00$, the classification performance is 100%, considering zero false positive detected objects at the same time. ROC curves are often used for *diagnosis* tasks because they can give a good description of the overall system performance, and they provide a means for evaluating two or more classifier models. Moreover, a plot of sensitivity versus false positives per image, is called a free-response receiver operating characteristic (FROC) plot, and this is generally used to report performance of the *detection* algorithm. In this study, we use AUC in our experiments for measuring the classifiers' performance.

5.11 Summary

In this chapter we described the theory behind machine learning, including supervised, unsupervised and reinforcement learning mechanisms. However, we focused and thoroughly described the fundamentals of *classification*, a form of supervised learning, in which algorithms are trained with labeled datasets that determine the class to which each instance belong to.

We presented the categories of classifiers that exist in the literature and then we further described those which are of interest for this research work: (1) neural networks, a popular and highly accurate algorithm that has been used in several domains; (2) support vector machines, a kernel-based method that also has been proved to be very effective in different classification problems; (2) the discriminant analysis method, a linear classifier that is built upon strong statistical theory and (3) the k-NN, which is a simple mechanism based on majority vote.

Also, we introduced important aspects of the Case-Based Reasoning (CBR) philosophy as a learning mechanism that solves problems by exploiting information that can be drawn from historical cases that are similar to the one in study. The CBR cycle includes four stages: *retrieving* similar cases from the knowledge base, *re-using* those examples in an appropriate way to come up with an answer and present them to an expert that *revises* the suggested output and who finally confirms or corrects the solution, before the new case is *retained* in the database.

Furthermore, in our study we also compared the proposed CBR solution model against machine learning algorithms that can be used as an alternative to build a solution. Specifically, we decided to use classifier ensembles with Bagging and Random Subspace, using the FURIA algorithm as the base classifier, and also combining them with techniques that are used to learn from imabalanced datasets (an endemic characteristic of almost any medical dataset and many other scenarios), such as SMOTE and the ENN rule. All the underlying processes of these techniques were also described in this chapter.

Chapter 6

Related Work

Breast lesions have a wide range of features that can indicate malignancy. However, they can also be part of benign changes in the breast and sometimes they cannot be distinguished from the surrounding tissue; these issues make the detection and diagnosis tasks even more difficult. Radiologist's misinterpretation of the lesion can lead to a greater number of false positives, since between 65% to 90% of the biopsies of suspected cancers turn out to be benign [21].

Hence, it is of major importance to develop methods capable of increasing the lesion-detection and cancer-diagnosis accuracy, for physicians to make better decisions regarding follow-up and biopsy. One powerful option is the use of computers in processing and analyzing biomedical images, such as mammograms, since they provide the radiologist with relevant information, which may not be readily observed by the physician, resulting in an enhanced detection/diagnosis process. Moreover, humans are susceptible to make mistakes, and their analysis is usually subjective and qualitative, while computers provide an objective and quantitative biomedical image analysis and, thus, leading the physician to a more accurate diagnostic decision [81].

Computer-aided detection (CADe) and computer-aided diagnosis (CADx) systems can improve the results of mammography screening programs and decrease the amount of false positive cases. Typically, in both CADe and CADx algorithms the following steps are performed: *preprocessing*, *feature extraction*, *feature selection* and *classification*. However, it should be noted that in some approaches, one or more of the previous steps may involve very simple methods or be skipped entirely. The output of the CADe process are the lesions detected and marked as regions of interest (ROIs); these ROIs can be masses, calcifications or other lesions which are being analyzed. The CADx phase take these ROIs, containing the abnormality, as input and determine the likelihood of malignancy or management recommendations as the output.

The *preprocessing* step has the objective of decreasing noise and improving the quality of the image. It is done on digital/digitized images; if a digital mammogram is given, no *digitizing* process is performed, but the preprocessing stage is still necessary. Another important task within this stage is to remove the background area and, if a MLO view is provided, to remove the pectoral muscle from the breast area. Furthermore, the *segmentation* step aims to find suspicious ROIs containing abnormalities and the *feature extraction* step computes the features of the given image, from the characteristics of the provided ROI. The *feature selection* step has the objective of selecting the set of features that are going to be used for eliminating false positives and for classifying lesion types; this stages is defined as selecting a smaller feature subset that leads to the largest value of some classifier performance function [49]. Finally, on the basis of selected features, the false positive reduction and lesion classification are performed in the *classification* step.

There are several research efforts aimed to developing novel techniques to deal with important

phases such as: **segmentation**, **feature extraction**, **feature selection** and **classification**. A description of very important, often cited, research efforts will be provided in the following sections.

6.1 Segmentation

The aim of segmentation step in mammographic image analysis is to extract regions of interest (ROIs) containing all breast abnormalities from the normal breast tissue. Another aim of the segmentation is to locate the suspicious lesion candidates from the region of interest.

Several techniques have been used to explore solutions in this stage. Dominguez and Nandi [28] performed segmentation of regions via conversion of images to binary images at multiple threshold levels. For images in the study, with gray values in the range [0, 1], 30 levels with step size of 0.025 were used to segment all mammographic images. Their method achieved a sensitivity of 80% and a 2.3 false-positives per image.

Singh et al. [87] propose a methodology for estimating the probability of each pixel in an image as being *suspicious* (i.e. belongs to a mass), based on the concept of using Gaussian mixture models, for modelling the data density function of pixels from the training data such that it can be represented as a weighted combination of more than one data density component; the weight coefficient for each Gaussian density component is determined using the expectation maximization (EM) algorithm. This research effort proposed a weighted Gaussian mixture model, for both supervised and unsupervised data analysis, with and without the addition of a Markov random fields hidden model, based on which the segmentation process is performed. The combination of different weighted Gaussian mixture models, each one trained on different image features, is shown to outperform the classic rule-based ensemble combination technique.

Li et al. [60] designed a method to automatically detect all possible sites in a mammography that may contain a suspicious mass in the breast tissue. In their work, they use a model-based image segmentation technique in which they propose one type of morphological operation to enhance the mass patterns on mammograms, considering the geometric features of this type of lesion against normal background tissue. Then, they use a finite generalized Gaussian mixture (FGGM) distribution to model the histogram of the mammograms where the statistical properties of the pixel images are largely unknown and are to be incorporated to their statistical model. Optimization techniques based on EM are also implemented to determine the optimal amount of image regions that should be considered and most suitable shape that should be used in the FGGM model. The final stage computes the selection of suspicious masses using a contextual Bayesian relaxation labelling (CBRL) method.

Several region growing methods were implemented by G.M. te Brake et al. [90] and compared against a discrete dynamic contour technique. They used a series of features related to the shape and intensity of the suspected mass region to discriminate between normal and abnormal regions and applied the considered methods to a dataset of 132 mammograms from the Dutch screening that contain masses and architectural distortions. The initial experiments compared the output of their segmentation system to annotations introduced by an expert radiologist and another set of experiments used features computed from all segmented areas, normal and abnormal, to classify then with a neural network and, thus, decrease false-positives.

Varela et al. [96] segmented suspicious regions using an adaptive threshold level. The image were previously enhanced with an iris filter at different scales. The output of the iris filter was used to characterize with features the suspected regions. False positive reduction was performed by using a trained backpropagation neural network, showing a sensitivity 88% and 94% at 1.02 false positives per image, considering the FROC curve.

Li et al. [61] used adaptive gray-level thresholding to obtain an initial segmentation of suspicious

regions followed by a multiresolution Markov Random Field Model-based method. The detection accuracy of this method was evaluated by using the FROC curve, reporting a 90% sensitivity in the detection of masses and 1.5 false alarms per image. They also evaluated their proposed method with minimal cancers manifested by masses less than 10 mm in size.

Wu et al. present in [83] one segmentation approach that is based on Reinforcement Learning (RL). In this approach, a learning agent takes specific actions, such as changing the tasks parameters, to modify the quality of the segmented image. The model starts with a limited number of training samples and improves its performance in the course of time. As expected, the reinforcement agent is provided with reward and punishment, which makes it explore and exploit the solution space.

The proposed system contains a series of image processing tasks, with parameters that must be adjusted to manipulate the input images in some way. Consequently, the goal is to choose a final set of parameters for several tasks, such that an object of interest is extracted.

The actions that the agent performs in its environment are modifications of the parameters of processing tasks; after the agent is done with an action, it receives a reward, which is provided by the environment and must accurately reflect the goal of the agent, typically by making use of an *evaluation metric*.

The authors of the method being described, present an object segmentation of two modes: offline and online. A general description of them is provided in the following:

1. **Offline Training:** it is conducted by using manually segmented samples. The agent is adopted in a simulation environment and interacts with training samples to provide information about the required parameters. Once the agent's actions are acceptable, it switches to the online mode.

2. **Online Mode:** the RL agent is used directly in a real-time case with new images; subjective feedback is provided in this mode. The agent is continuously learning and therefore can adapt to changes in the input images.

As for the input images, they are divided into $R_I \times C_I$ subimages (for R_I rows and C_I columns), and the agent works on each of them separately. Local processing on subimages is conducted to find the segmentation parameters for each of them, and finally a processing task chain is constructed.

6.2 Feature extraction

Once the segmentation phase has found suspicious ROIs containing abnormalities, the feature extraction phase is performed, in order to calculate features from the characteristics of those regions of interest. Regarding mammographic images, there are several features that are taken into account, depending on the objectives of the detection process. A list of research efforts aimed at computing features from mammographic images is provided in this section and, also, some of them will be briefly described.

For instance, in [92], Timp et al. investigated the effect of temporal features on the performance of mass detection, considering two consecutive mammographic screening rounds, in which they implemented a pixel level mass detection algorithm which links a suspicious location on the current mammogram with a corresponding location on the prior mammogram; to achieve that goal, two features were calculated for each pixel to detect stellate lesions and two features to detect lesions with a focal mass. Later, a neural network classifier trained on those characteristics assigned each pixel a measure of suspiciousness using a dataset of 589 cases.

Hadjiiski et al. [33] developed an approach to classify mammographic masses as malignant or benign, by using interval change information. For segmentation purposes, in both current and prior

mammograms, they used an active contour method, and from each resulting ROI, 20 run length statistics (RLS) texture features, 3 speculation features and 12 morphological features were extracted. An additional 20 difference RLS features were obtained by substracting the prior RLS features from the corresponding current RLS features. They showed that taking into account prior mammographic images improves the accuracy for classifying masses.

Verma et al [98] claimed that the combination of three features (e.g. entropy, standard deviation and number of pixels) is the best strategy to distinguish a benign microcalcification pattern from a malignant one. They took into account 14 features, which were combined and applied to a neural network, in order to find out what was the most appropriate combination; the features used in this research work are: average histogram, average gray level, energy, modified energy, entropy, modified entropy, number of pixels, standard deviation, modified standard deviation, skew, modified skew, average boundary gray level, difference and contrast.

The performance of using textural features to detect breast masses on mammograms is explored by Choi et al. [53]. They designed a method for extracting local binary pattern (LBP) textural features from identified mass-regions with the objective of reducing the false-positive detection rate. With these features they are able to characterize texture patterns from important regions of a suspicious mass, such as the margin and the core; also, the extracted features preserve the spatial structure information of the mass. They conducted comparative experiments to evaluate the proposed model on two mammogram datasets.

Features computed based on gray-level co-occurrence matrices (GCMs) are used by Mudigonda et al. [73] to evaluate the performance of textural information possessed by mass regions in comparison with the textural information present in mass margins. They proposed a method that involves polygonal modeling of boundaries for the extraction of a ribbon of pixels across mass margins and two gradient-based features were developed to estimate the sharpness of mass boundaries in the ribbons of pixels extracted from their margins. The Mammographic Image Analysis Society (MIAS) database was used for the evaluation of their model, with a total of 54 images representing 28 benign and 26 malignant cases.

6.3 Feature selection

This section provides a description of the *feature selection* process, along with a list of relevant research works on this topic; once again, only some of them will be described in the following.

As previously mentioned, typically there are several mammographic features that can be taken into account, in order to eliminate false positives, i.e. regions of interest erroneously identified as lesions. Furthermore, some of those regions of interest are not significant when observed alone, but when combined with other features, they can be significant for classifying lesion types as benign or malignant; the set of features that leads to eliminating false positives for classifying suspected lesions is selected in the feature selection step.

Several techniques have been implemented, in order to perform this feature selection process. A statistical feature selection method was proposed by [32], in which they keep features that are statistically significant via a t-test of 95% confidence level. Then, they used a wrapper feature selection approach considering the performance of a SVM classifier to detect relevant subset candidates and, finally, they implemented a filter method to eliminate those features whose contribution to classification accuracy was marginal. The output of the proposed model is a compact yet highly diagnostic feature subset.

Hernández-Cisneros et al [40] used a GA for selecting the most relevant features extracted from individual microcalcifications and from microcalcification clusters, that later on the process become

the inputs of two simple feedforward neural networks for the classification of microcalcifications and microcalcification clusters in digital mammograms. They compared the use of GAs with the forward sequential search presented in [75], in which inputs were sequentially added to the neural network while its error decreases, and stopping when it starts to increase. They showed that the use of GAs and neural networks greatly improves the overall accuracy, specificity and sensitivity of the classification. The best solution found was a neural network with 23 inputs, corresponding to 23 extracted features and, in average, they used 20 features (less than the half of the extracted features).

They also provided a more extensive study in [41] and [24], comparing the GA approach for selecting the most relevant features extracted from both individual microcalcifications and microcalcification clusters, against the methods used in [15] and [16], in which features are ordered according to their class separability and then selecting the most relevant ones and also against the method used in [75], which is based on a forward sequential search. Once again, the proposed method needed less than the half of the original 47 features to improve the overall accuracy, specificity and sensitivity of the classification; it was observed that the class separability-based method had similar results in the case of specificity and good overall performance, but had poor accuracy regarding the specificity of the classification process.

6.4 Classification

By using the features previously extracted and selected, the classification step in the mammographic image analysis, identified breast lesions are classified as benign or malignant.

Casti et al. [76] proposed the use of two novel feature descriptors based on 2D extensions of the reverse arrangement (RA). They computed the radial correlation and radial trend from the original gray-scale values as well as from the Gabor magnitude response of 146 regions of interest, of which 120 were benign and 26 malignant tumors. Finally, they employed Linear Discrminant Analysis (LDA), a Bayesian classifier, a support vector machine (SVM) and an artificial neural network with a radia basis function, to predict the malignancy of each suspicious region, considering also a stepwise logistic regression for feature selection and a leave-one-patient-out approach for cross-validation.

Oporto-Díaz et al [75] proposed a method for detection of microcalcification clusters in mammograms using sequential difference of gaussian filters. In the first stage, fifteen of those filters are applied sequentially to extract the potential regions, and in a later phase, these regions are classified using absolute contrast, standard deviation of the gray level of the microcalcification and a moment of contour sequence. In this stage, the performance achieved for classifying signals into microcalcifications was reported to be 70.8% for true-positives and 85.7%. Then, two strategies for clustering methods are compared: (1) considering several microcalcification clusters detected in a given mammogram, and (2) considering all microcalcifications as being in a single cluster. This research work showed that the diagnosis based on the detection of several microcalcification clusters in a mammogram results in better performance, than considering always all microcalcifications present in a given image as a single and unique cluster. The performance achieved at diagnosing a microcalcification cluster was reported to be 91%.

In [38, 39], Hernńdez-Cisneros et al propose a procedure for the classification of microcalcification clusters in mammograms using, once again, the sequential difference of gaussian filters, and three evolutionary artificial neural networks, compared against a feedforward artificial neural network trained with backpropagation. The model uses genetic algorithms for three purposes: finding the optimal weight set for an artificial neural network, finding an adequate initial weight set before starting a backpropagation training algorithm and designing its architecture and tuning its parameters. All those methods are applied to the classification of microcalcifications and clusters of microcalcifications in

digital mammograms. Results show that the implemented evolutionary method is better than the simple backpropagation method for the classification of individual microcalcifications, regarding specificity and overall accuracy; however, in terms of sensitivity, backpropagation was significantly better than only evolving weights. Furthermore, in the case of classifying microcalcification clusters, the performance of evolutionary methods is significantly better than that of the simple backpropagation. The method that evolves initial weights, complemented with backpropagation, is the one that gives the best results.

A more extensive study regarding the benefits of using genetic algorithms, instead of simple backpropagation methods, in order to classify microcalcifications clusters in mammograms, is provided in [42]. Again, improvements regarding overall accuracy, sensitivity and specificity of the selected classifier (an artificial neural network) are achieved.

According to Wu et al. [102], neural networks (NN) have been incorporated into many CAD systems, typically distinguishing cancerous signs from normal tissues. In a general basis, the strategy to follow is to enhance the images first, in order to extract the regions of interest from them; this is important to do, since many features are calculated based on the extracted regions and are then forwarded to NN, to make decisions. The most typical examples of these applications are early diagnosis of breast cancer and lung cancer.

Ge et al. [30] developed a CAD system to identify microcalcification clusters automatically on full-field digital mammograms. Main stages of this system included the following six stages:

1. Preprocessing.
2. Image enhancement.
3. Segmentation of microcalcification candidates.
4. False Positive reduction for individual microcalcifications.
5. Regional Clustering.
6. False Positive reduction for clustered microcalcifications.

To reduce false positive individual microcalcifications, a convolution NN was employed to analyze 16×16 regions of interest centered at the candidate derived from segmentations. This NN contained an input layer with 14 neurons, two hidden layers with 10 neurons each, and one output layer. The convolution kernel sizes of the first group of filters between the input and the first hidden layer were designed as 5×5 and those of the second group of filters between the first and second hidden layers were 7×7.

The images in each layer were convolved with convolution kernels to obtain the pixel values to be transferred to the following layer; the logistic sigmoid function was chosen as the transfer function for the hidden and output neurons.

The convolution NN was trained using a backpropagation learning rule with the sum-of-squares error (SSE) function; this provided the probability of correctly classifying the input sample as a true microcalcification region of interest. At the stage of false positives reduction for clustered microcalcifications, some morphological features (size, mean density, etcetera) along with some convolution NN outputs (such as the minimum, maximum and mean of output values) were extracted from each cluster. They used a set of 96 images, which were split into a training set and a validation set, each with 48 images.

On the other hand, medical diagnosis systems based on a CBR methodology have been developed for different domains. Armengol [7], proposed a method for classifying melanomas *in situ* using

clustering techniques under the CBR philosophy. In this work, the clustering algorithm is not used to organize the case base to enable an efficient retrieval mechanism; rather, they proposed a method in which CBR is used for clustering in which a lazy learning method produces *explanations* that are used as clusters' descriptors. This way, clusters can show the expert a picture of some parts of the domain of the problem, with its main characteristics.

A biofeedback training method proposed by Ahmed et al. [3] uses modified distance function, similarity matrix and fuzzy similarity for case-retrieval. The system receives a time series signal related to finger temperature and uses a CBR approach to support classification of patients, parameter estimation and biofeedback training in stress medicine and can be queried by less experienced physicians who can also use this system as a decision support tool which provides a second opinion in diagnosis tasks. In every module of this system, the CBR technique retrieves the most similar cases from the knowledge base by comparing a new finger temperature reading with previously solved measurements.

A cost-sensitive learning approach was presented by Park et al. [78], incorporating unequal penalizations for misclassifying positive or negative cases, all within a CBR model. This research work describes a method based on a GA to find the optimal cut-off point to discriminate between malignant and benign cases, as well as the cut-off point for determining the optimal number of neighbors that should be retrieved from the case base, considering their similarity scores. This cost-sensitive method allows to take into account that false-negatives have more severe implications than false-negatives and that, therefore, they should receive different penalization costs.

Furthermore, databases are typically indexed to enable an efficient search for case retrieval. Common indexing techniques include kd-tree [8] and Locality Sensitive Hashing (LSH) [31, 25, 6]. The former creates a partition of the feature space using a tree structure; the latter uses hash functions to compute the distance between the query and a subset of reference points.

6.5 Summary

In this chapter we introduced several related research efforts that also propose a solution for breast cancer computer-aided diagnosis. There are several stages that have to be considered in designing solutions for this domain, such as: lesion segmentation, feature extraction, feature selection and classification; we described scientific papers that have addressed one or more of these stages.

Chapter 7

Solution Model

Having all the previous technical facts and theory as background, this chapter introduces the proposed model that was implemented as a solution for breast cancer computer-aided detection (CAD). As the reader may recall, there are several modules that are involved in the design of a CAD system, going from pre-processing the input image all the way to automatically detect and diagnose suspicious regions; this chapter provides a description of all the stages that we designed for our model.

The processes that the proposed solution model include can be grouped in two separate frameworks: (1) *database building* and (2) *classification* frameworks; each of them is made up of several processes, as depicted in Figure 7.1. A description of them is provided in the following sections.

The reader will realize that the model is designed based on the *Case-based Reasoning* (CBR) approach, which was described previously in section 5.6, since all key activities from the CBR philosophy are implemented. The input mammogram is processed here as a new case, that determines the input query that is used to ***retrieve*** similar cases from the database; then, a classification algorithm ***reuses*** this data to provide a solution to the new case, which is ***revised*** by the user. At the end of this process, the revised solution is ***retained*** in the case base.

7.1 Database Building Framework

The objective here is to design and implement a method for building and managing a database of **mammographic images** and their **meta-data**. This is an off-line process, which means that it is not necessarily performed in parallel with the *classification framework*. Rather, it has to be performed at the very beginning of the project, considering all mammographic images available at that time, in order to enable the CBIR tasks conducted to retrieve images and the consequent classification of breast lesions. Afterwards, new mammograms can always be analyzed and stored in the database at any given time, without negatively affecting the *classification framework*; on the contrary, adding more cases to the database should increase the number of mammograms that can be retrieved from it and, hence, the classifier will have more information about previous experiences to take advantage from, as mentioned in section 5.6.

Following the CBR approach, the user's *revised solution* will be *retained* in the case base, for future usage. Within this **Retain case** process, different database management activities can be performed, considering the modification of existing cases, if necessary, or simply storing the new case in the repository.

It is worth mentioning that, as depicted in Figure 7.1, both frameworks perform some common processes; this provides the system with homogeneity between its entities, hence, preventing the design itself from adding external noise to the problem which is being solved. Since these processes comprise

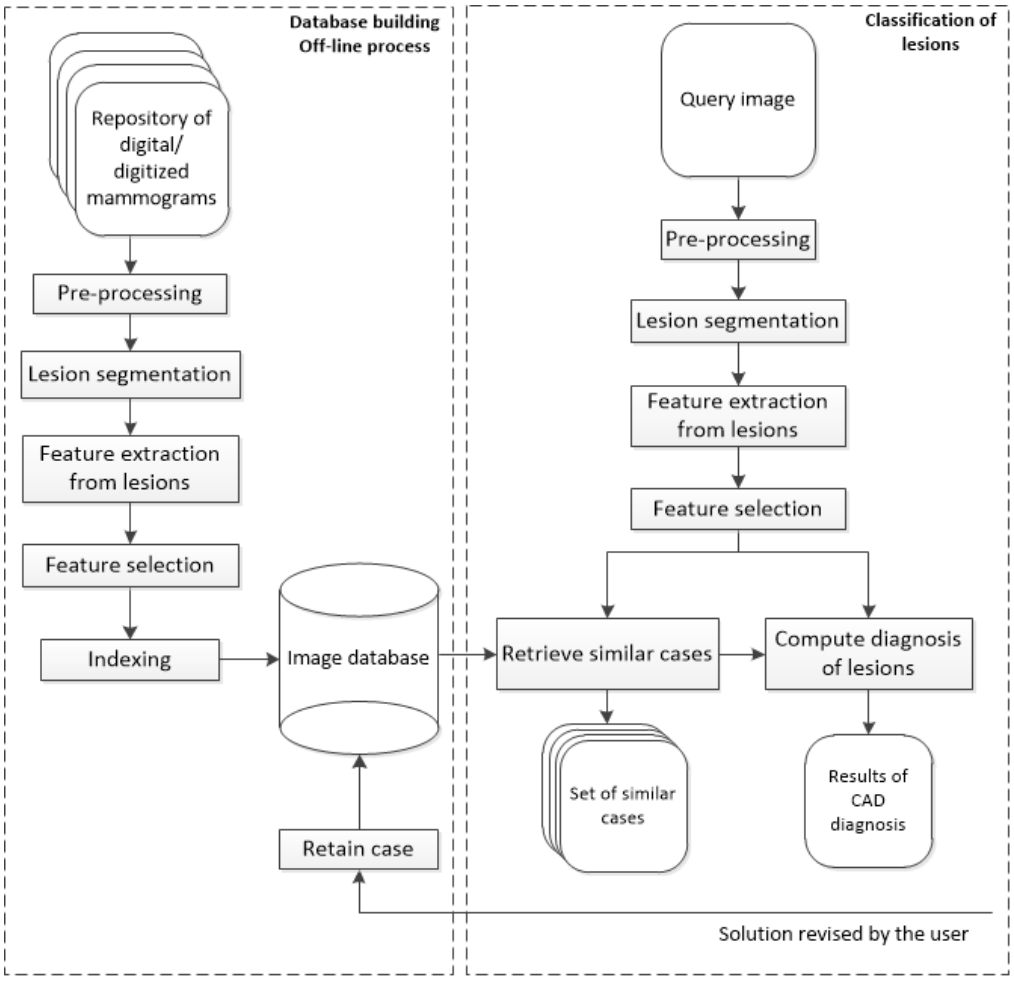


Figure 7.1: Proposed Solution Model

the exact same underlying activities, they will only be described once, during the definition of the *Classification Framework*.

7.2 Classification Framework

This framework comprises the processes related to the analysis of a mammographic image in order to detect and classify MCC and masses into benign and malignant; additionally, the retrieval of cases similar to the one being currently analyzed is also conducted in this framework, with the objective of using that information to enhance the performance of a classification algorithm and, also, displaying images to the end-user (the radiologist). A thorough description of the processes performed in this framework is provided in the following sections.

7.2.1 Pre-processing

Mammograms often present noise and artifacts that were acquired during the creation of the x-ray image and/or during the digitization of a hard mammography, due to changes in illumination and distortions in the properties of the digitized image. Besides, the background of a typical mammogram contains labels and marks that are not useful for the analysis of breast cancer itself; and therefore have to be removed, to prevent the subsequent stages from being negatively affected.

This pre-processing method takes the original mammography as input and applies a median filter to eliminate noise located in the background, while keeping important features of the image, like the breast tissue. In this process, a 3×3 mask was used, centered in each pixel within the image, and the value of that pixel was replaced by the median of the surrounding mask pixels. The size of this mask was chosen empirically, trying to avoid the loss of local details, as described in [24]. Furthermore, the background marks and the isolated regions are deleted, for the image to contain only the breast tissue, by way of a binarization process in which all pixels that are not within the group of those corresponding to the breast are removed. The complete process is depicted in Figure 7.2.

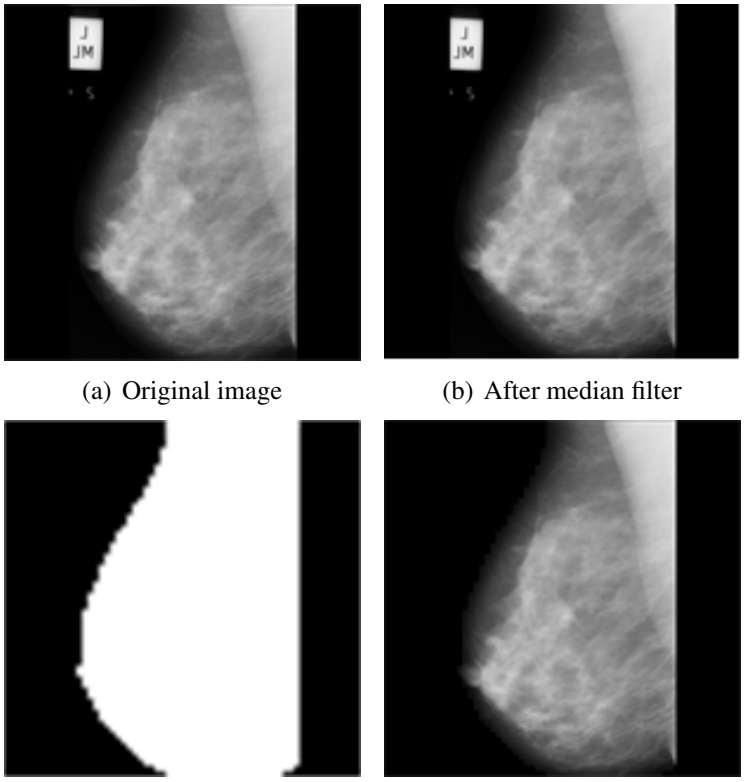
7.2.2 Detection of Lesions

Having the pre-processed image as an input, the objective of this process is to automatically detect regions of the mammogram containing potential breast microcalcification clusters and masses. The methods described here were previously published by the author and members of his research group in [5].

Figure 7.3 depicts the stages related to the process of detecting clusters of microcalcifications, applied to a representative case, shown in Figure 7.3(a). The first step is to highlight microcalcifications, by way of contrast adjustment, followed by a negative filter and, lastly, a mean filter with a 2-pixel window since calcifications are usually represented by small regions.

Next, the resulting image is used to find the edges of microcalcifications, by applying an edge detector based on a Difference of Gaussian (DoG) filter with a 26-pixel window and a theta value of 0.363. The resulting image is shown in Figure 7.3(b). Then, we execute a Gaussian Blur filter with a 3-pixel window and a theta value of 4.8 to smooth the image for further processing.

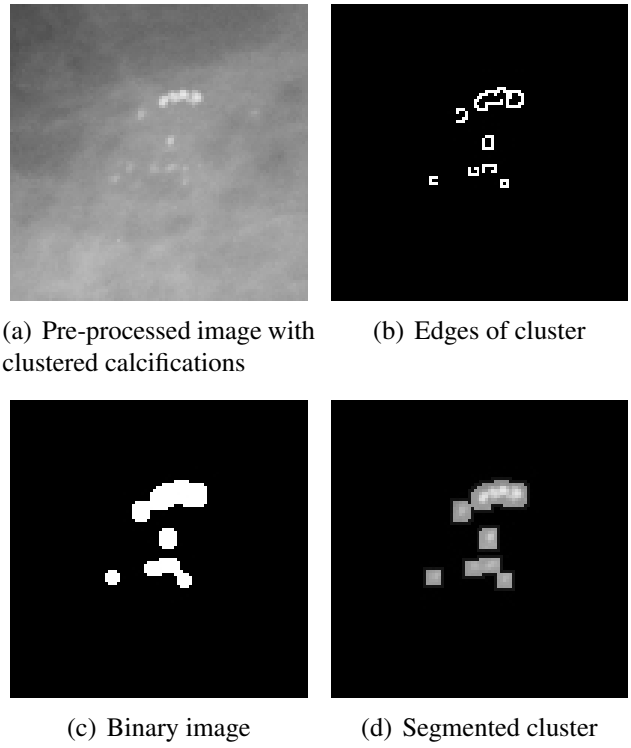
The third step consists of binarizing the image, using a gray-scale threshold that narrows the complete scale (0-255) to a more representative 25-230 scale, resulting in the image shown Figure 7.3(c). Finally, we use the binarized image as a mask to segment the original image and extract the regions of interest, providing the image presented in Figure 7.3(d) as a result.



(a) Original image (b) After median filter

(c) Binary image without background labels (d) Pre-processed image

Figure 7.2: Steps of the pre-processing stage, using image mdb219 from MIAS database.



(a) Pre-processed image with clustered calcifications

(b) Edges of cluster

(c) Binary image

(d) Segmented cluster

Figure 7.3: Steps in the detection of microcalcifications stage.

On the other hand, the detection and segmentation of masses was performed by using seven procedures: (1) location of masses using an adaptive global threshold, (2) edge detection, (3) region-growing edges, (4) elimination of overlaps, (5) union of edges, (6) internal binarization and (7) segmentation. Figure 7.4 illustrates the stages of this process.

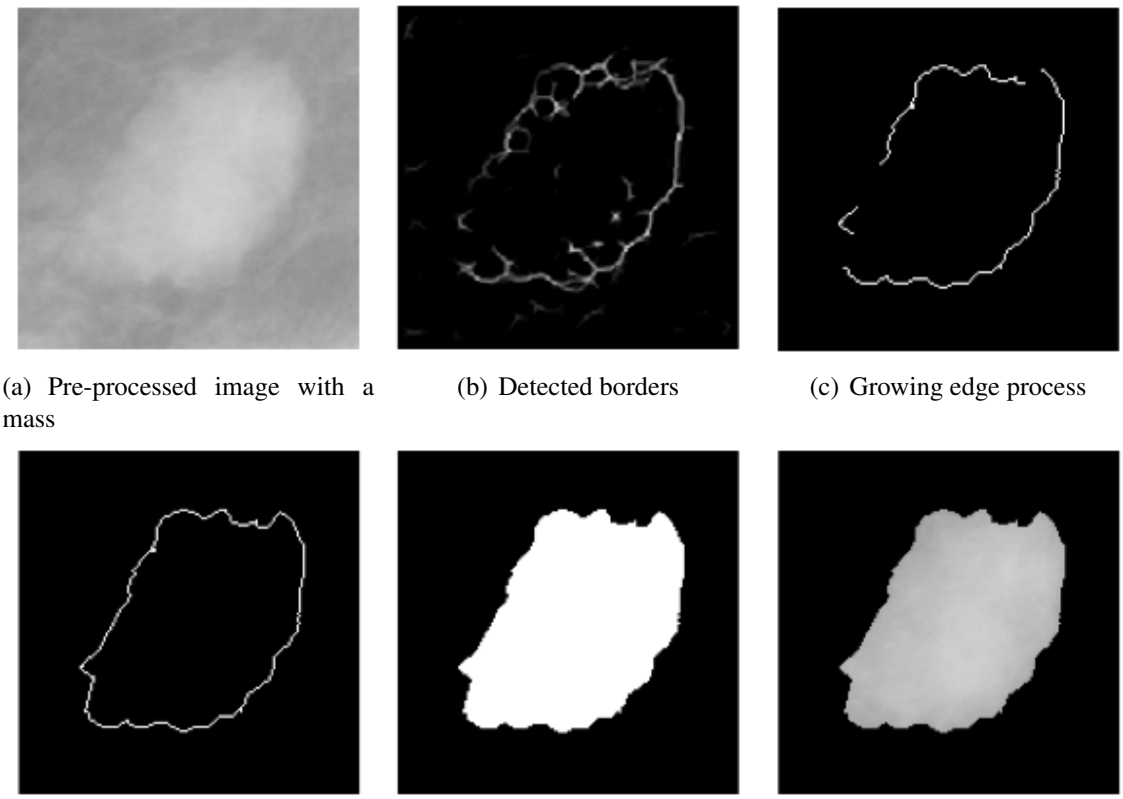
In order to determine the location of suspicious masses, we use a global adaptive threshold segmentation process, which analyzes the global information of the image based on its histogram and determines an appropriate threshold to segment different areas [27]. At this point the masses have been located in the image and the following procedures will focus only on those regions containing masses.

Once the masses have been located, we proceed to detect its edges, in order to determine its shape and contour, with the objective of obtaining necessary information for the following phases. The technique used for this purpose is edge detection based on wavelet transform. This edge detection algorithm accumulates multiscale-wavelet edges and generates an image with some points that do not necessarily represent the margin, as depicted in Figure 7.4(b). Thus, a refinement process should be conducted.

Consequently, in order to refine the line of the margin of the mass, we use a region-growing edge technique, that starts from some seed points and it continues by adding representative pixels iteratively verifying the neighbors of pixels that meet certain criteria [10]. Figure 7.4(c) shows the resulting image up to this point.

The next step to refine the margin of the mass was the elimination of overlaps. To accomplish this process we drew radial lines from the centroid of the region to each end of each edge. If one those radial lines touched two or more different edges, an overlap had been found. In those regions where two adjacent radial lines touched more than one edge, the area between those lines was binarized and used as mask to eliminate that overlap. This process was applied to every edge of the mass. As result of this elimination of overlaps process, the edges of the mass have been located as unconnected lines.

Once the edges of the masses have been identified, we perform a process aimed to connect the edges that were found. This union process draws straight lines from the end of one edge to the start of its nearest one, as depicted in Figure 7.4(d).



(a) Pre-processed image with a mass (b) Detected borders (c) Growing edge process

(d) Connection of edges (e) Binarized image (f) Segmented mass

Figure 7.4: Steps of the mass detection stage.

Once we have a continuous line as the border of the mass we run the binarization process, which results in the image presented in Figure 7.4(e). The purpose of the binarization of the internal region surrounded by the edge is to have a binary image that will serve as a mask for cutting the area comprising the mass in the original image. Finally, the segmentation procedure extracts the area corresponding to the mass to perform subsequent calculations over this area, as depicted in Figure 7.4(f).

7.2.3 Feature Extraction

Once the lesions were detected and extracted as ROIs, they are used to compute a set of features that describe their visual content, with the objective of enabling a classification algorithm to discriminate between benign and malignant cases.

Having detected all individual calcifications in one image, an algorithm for locating regions where density (the amount of them per cm^2) is higher was performed, in order to detect so-called clusters of microcalcifications. For this, all pairs of individual microcalcifications within an Euclidean distance less than an empirically defined threshold of 100 pixels are regarded to be a subset of *neighboring microcalcifications*; these subsets are explored to find the one with the highest density and select it as a new cluster. The process is then repeated with the remaining ungrouped microcalcifications until all are included in a labeled cluster.

Afterwards, all detected clusters are passed on to the *feature extraction* process, in which 31 features, listed in Table 7.1, are computed from them. There are 14 features which define the shape of a cluster, 6 which describe the area of the individual microcalcifications and 11 for the absolute contrast between the calcifications and their background.

	Computed Features
Cluster shape (14 features)	Number of calcifications, convex perimeter, convex area, compactness, microcalcification density, total radius, maximum radius, minimum radius, mean radius, standard deviation of radii, maximum diameter, minimum diameter, mean of the distances between microcalcifications, standard deviation of the distances between microcalcifications.
Area of Microcalcifications (6 features)	Total area of microcalcifications, mean area of microcalcifications, standard deviation of the area of microcalcifications, maximum area of the microcalcifications, minimum area of the microcalcifications, relative area.
Microcalcification Contrast (11 features)	Total gray mean level of microcalcifications, mean of the mean gray levels of microcalcifications, standard deviation of the mean gray levels of microcalcifications, maximum mean gray level of microcalcifications, minimum mean gray level of microcalcifications, median of the man tray level of microcalcifications, total absolute contrast, mean absolute contrast, standard deviation of the absolute contrast, maximum absolute contrast, minimum absolute contrast.

Table 7.1: Features extracted from microcalcification Clusters

Also, once all masses have been detected in a mammogram, they are passed on to the feature extraction process, in which 50 features are computed from them. The complete list of features for masses is shown in Table7.2. There are 7 features that describe the signal contrast, 7 features that define background contrast, 3 features for representing the relative contrast of the image and 20 features for shape. There also are 6 features for contour sequence moments and 7 features that describe the first invariant moments.

7.2.4 Feature Selection

The next step of the analysis consists in taking the set of features extracted for each ROI into a feature selection process. The main purpose of this phase is to find the subset of the whole set of features that most contribute to the performance of a given classifier and, also, has a reduced dimensionality.

This process was carried out using a wrapper approach based on a GA and four different classifiers: the k-nearest neighbors (k-NN) [35], a feed-forward back-propagation neural network (NN) [36], support vector machine (SVM) classifier [95] and the linear discriminant analysis (LDA) method [35, 34].

The implemented wrapper used a GA as search algorithm to explore through the space of possible feature subsets, taking advantage from its ability to exploit accumulating information about an initially unknown search space in order to bias subsequent search into promising subspaces [52]. It works with

	Computed Features
Signal contrast (7 features)	Maximum gray level, Minimum gray level, Median gray level, Mean gray level, Standard deviation of the gray level, Gray level asymmetry (skewness), Kurtosis of gray level.
Background contrast (7 features)	Maximum gray level, Minimum gray level, Median gray level, Mean gray level, Standard deviation of the gray level, Gray level asymmetry (skewness), Kurtosis of gray level.
Relative Contrast (3 features)	Absolute contrast, Relative contrast, Proportional contrast.
Shape (20 features)	Area, convex area, background area, filled area, perimeter, maximum diameter, minimum diameter, orientation, eccentricity, Euler number, circular diameter equivalent, solidity, Extent, shape factor, roundness, aspect ratio, elongation, compactness 1, compactness 2, compactness 3.
Contour sequence moment (6 features)	Contour sequence moment 1, contour sequence moment 2, contour sequence moment 3, contour sequence moment 4, mean radii, standard deviation of radii.
First invariant moments (7 features)	Invariant moment 1, invariant moment 2, invariant moment 3, invariant moment 4, invariant moment 5, invariant moment 6, invariant moment 7.

Table 7.2: Features extracted from Masses

a population of candidate solutions, or individuals, and through the generations looks for the fittest individual.

In our feature selection process each individual of the population represent a subset of features. The chromosomes of the individuals in the GA contain an amount of bits equal to the total number of features, i.e. one bit for each extracted feature. Consequently, the chromosomes of the individuals in this study had lengths of 30 for clusters of microcalcifications and 50 for masses, corresponding to the amount of features extracted for each kind of lesion.

The value of each bit within the chromosome determines whether feature will be selected or not, and therefore if it will be considered in the subset of features provided as input to the classifier. Figure 7.5 shows an individual that determines that features 1, 3 and n were selected to be part of the classifier input, while feature 2 was not.

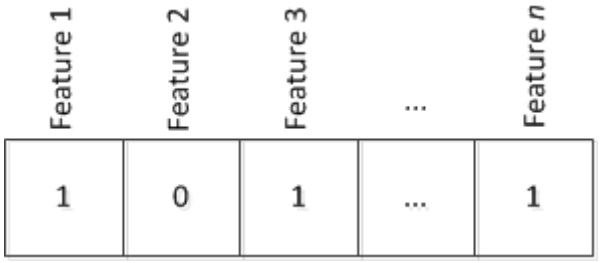


Figure 7.5: A sample individual of the GA which determines that Feature 1, Feature 3 and Feature n are selected.

The evaluation process by which the GA computes the fitness of each individual is depicted in

Figure 7.6. For any given query case, the subset of features determined by each individual is taken to perform a k-nearest neighbors similarity search through database of historical cases. The result of that search is a set of k cases that are similar to the query case and which are used to construct and train a classifier algorithm. This process is performed individually for the four classification methods considered in this study.

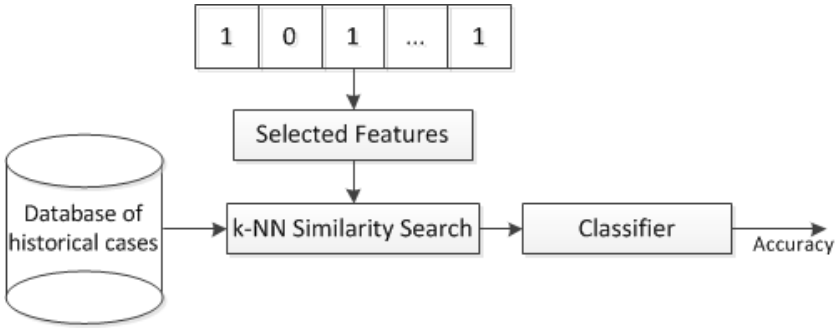


Figure 7.6: Computing the fitness of individuals in GA-based Feature Selection

Afterwards, a leave-one-out cross-validation through the whole database of historical cases is performed to compute the individual's fitness, which is determined by the AUC that a given classifier achieves if it uses the subset of features determined by the individual. AUC can be approximated with equation 5.34 presented in section 5.10.

The GA can stop due to two reasons: either the generations' limit or the maximum number of evaluations of the fitness function has been reached. The best individual of this evolutionary process determines the subset of features that have the highest discriminant power with respect to a given classifier.

7.2.5 Retrieving Similar Cases and Reusing them to Compute Diagnosis of Lesions

Retrieving and reusing similar cases are key processes within the CBR paradigm, since they enable solving new problems by looking at relevant information that can be drawn from the solutions of similar cases that were revised and validated by an expert sometime in the past. In this section we will explain how these processes are conducted within our model.

Our similarity search is computed upon the feature-vector that contains only the most discriminant ones, which were selected by the GA-based feature selection method. Therefore, it should be noted that in this work, *feature selection* is not only of major importance for optimizing the classification accuracy of the model, but it also enables a more efficient similarity search in the database, due to the associated dimensionality reduction.

The database of our model contains a set of mammographic images and the feature vector of the lesion that was found in each one. It is indexed by a kd-tree upon which the retrieval of cases is performed in a *k-Nearest Neighbors Similarity Search* paradigm, as depicted in Figure 7.7, with the objective of optimizing the retrieval process. It can also be observed that the activities involved in this stage are grouped in two different procedures, corresponding to the retrieval of similar cases and the computing the diagnosis of the new case, by training classifiers with the retrieved instances.

As we depict in Algorithm 3, in order to retrieve similar cases, the *dissimilarity score* between the new case and those stored in the database is computed by applying a distance metric to the feature vector of the query case $\mathbf{x}$ and any feature vector stored in the database. In this study, we explore

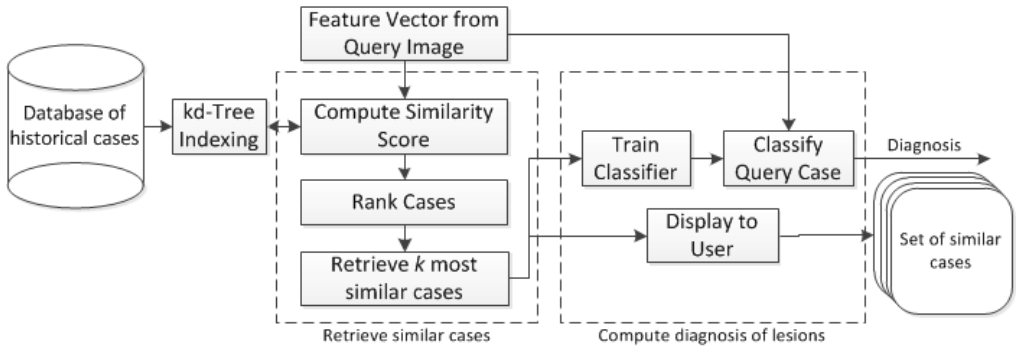


Figure 7.7: Retrieving similar cases and reusing them for classification of lesions

the performance of six different dissimilarity metrics: euclidean distance, manhattan distance, linear correlation, cosine similarity, Mahalanobis distance and the Spearman distance. We measured the AUC obtained by a majority-vote k-NN classifier that uses the considered dissimilarity metrics and conducted pairwise statistical tests between them to determine the one that is more suitable for the retrieval of cases.

Algorithm 3 Retrieving and reusing similar cases

Require: x, the feature vector of the query case; s, the similarity metric; k, the number of similar cases to be considered; Idx the kd-tree index of the database of historical cases and c, the classifier to be used.

procedure RETRIEVESIMILARCASES(x, s, k, Idx)
  for all $case \in Idx$ **do**
    $case.similarity \leftarrow computeSimilarity(x, s)$
  end for
  $rankedCases \leftarrow rankBySimilarity(Idx)$
  $kMostSimilar \leftarrow selectTopK(rankedCases, k)$
  $ReuseSimilarCases(kMostSimilar, x, c)$
end procedure

Require: T, the set of k most similar cases retrieved from the data base; x, the feature vector of the query case and c, the classifier to be used.

procedure REUSESIMILARCASES(T, x, c)
  $classifier \leftarrow train(T, c)$
  $Dx \leftarrow test(classifier, x)$
  return Dx
end procedure

Then, our methodology ranks the set of stored instances in descending order, based on the the similarity score that was previously computed between the query case and the historical ones and, finally, the k most similar ones are reused in computing the diagnosis of detected breast cancer lesions.

This assessment of malignancy is performed by taking the k retrieved instances as the training set of a given classifier algorithm. In this way, the classifier is going to be fed a set of relevant historical cases that contain data from lesions that have similar features to those encountered in the query case,

focusing the learning process in a subset of training samples within a neighbourhood of similarity instead of the whole dataset. This training process is executed for every new query case, providing the classifiers with the ability to adapt to the new problem, by re-using each time the most relevant instances to solve it.

To compute the diagnosis of lesions we are considering four classifiers: k-NN, NN, SVM and LDA, as we previously mentioned. Each of them are tested separately within our CBR framework with the aforementioned feature selection and training strategy. The performance measures that are considered for evaluating classifiers are the AUC, overall accuracy, sensitivity, and specificity (see equations 5.34, 5.26, 5.27 and 5.28, in section 5.10).

Moreover, it is worth mentioning that our CBR model is not dependant of specific classification algorithms, but rather describes a series of processes that are aimed to the classification of breast cancer cases in which learning algorithms are integrated as a tool to discriminate between benign and malignant lesions.

In this research effort we have integrated four different types of classifier algorithms, so as to explore the performance that can be achieved within the proposed CBR framework, from different learning perspectives. The NN is a robust, *non-linear* classifier that is highly adaptable and has been tested in several classification domains; the SVM represents a *kernel-based* classifier that has strong theoretical foundations and a good generalization capability and has recently become very popular as a high-performance classifier in several domains, as well. A *linear* approach is explored by integrating the LDA and a simplistic *majority-vote* strategy is represented by the k-NN algorithm.

7.3 Summary

The proposed solution model and its underlying processes were thoroughly described in this chapter. It has been designed under the Case-Based Reasoning (CBR) learning strategy and the main idea is to dynamically train classification algorithms with a set of similar cases that are retrieved from a database that contains instances with a validated diagnosis.

At the beginning of the method, a pre-processing stage is performed with the objective of deleting noise and removing all elements of the image that do not belong to the breast tissue. Afterwards, we perform a series of computer vision techniques to segment microcalcification clusters and masses from the mammographic images. Upon those potential lesions, a set of features that characterize their visual content is extracted and, in order to find the subset of features that are most discriminative, we perform a *wrapper* feature selection mechanism based on a GA that uses the AUC of the classifier as fitness function.

A k-NN similarity search is then conducted by applying a given similarity metric on the vector of selected features that were previously extracted from the query image and those stored in the database. Afterwards, the elements from the database are accordingly ranked and the k examples which are found to be most similar to the query case are retrieved and used for training a given classification algorithm and assessing the malignancy of the query case.

Chapter 8

Performance of Classifiers Under a Traditional CAD Pipeline

In this chapter we present the experiments that measure the performance of the considered classifier algorithms without being under the proposed CBR methodology; this can be regarded as the *traditional* pipeline of a CAD system, composed of a series of *sequential* processes in which classifiers are trained in advanced and remain static throughout the testing phase. Refer to Figure 1.1 for a complete description of the traditional CAD pipeline.

8.1 Experimentation Setup and Methodology

Figure 8.1 depicts the *traditional* approach that several CADs systems use for the classification of potential lesions. We have considered all the methods that were previously described, aimed at the automatic detection of microcalcification clusters and masses in digital mammograms that are fed as input, to implement this CAD approach and conduct this set of experiments. Feature selection and the classification of cases is performed using the feature sets described in Table 7.1 and Table 7.2, which are extracted from the encountered ROIs.

In the first stage we implemented the *pre-processing* mechanism, which has the objective of eliminating the elements in the mammographic image that could negatively affect the subsequent processes of detecting potential lesions, both calcifications and masses. Then, the detection processes are performed separately and once the lesions are segmented and extracted as ROIs, a set of features are computed in order to characterize their visual content. Then, using a GA, our system determines the subset of features that provides the highest classification performance, which is finally used in the classification step to determine whether the detected lesion is considered a malignant case or not.

We determined to integrate four different types of classifier algorithms, so as to explore the performance that can be achieved within the proposed CBR framework, from different learning perspectives. A neural network (NN) [36] is a robust, *non-linear* classifier that is highly adaptable and has been tested in several classification domains; the support vector machine (SVM) classifier [95] represents a *kernel-based* classifier that has strong theoretical foundations and a good generalization capability and has recently become very popular as a high-performance classifier in several domains, as well. A *linear* approach is explored by integrating the linear discriminant analysis (LDA) method [35, 34] and a simplistic *majority-vote* strategy is represented by the k-nearest neighbors (k-NN) [35] algorithm.

As we mentioned previously, the database of digital mammographies used in this research work was provided by the Mammographic Image Analysis Society (MIAS) [88], out of which we detected a

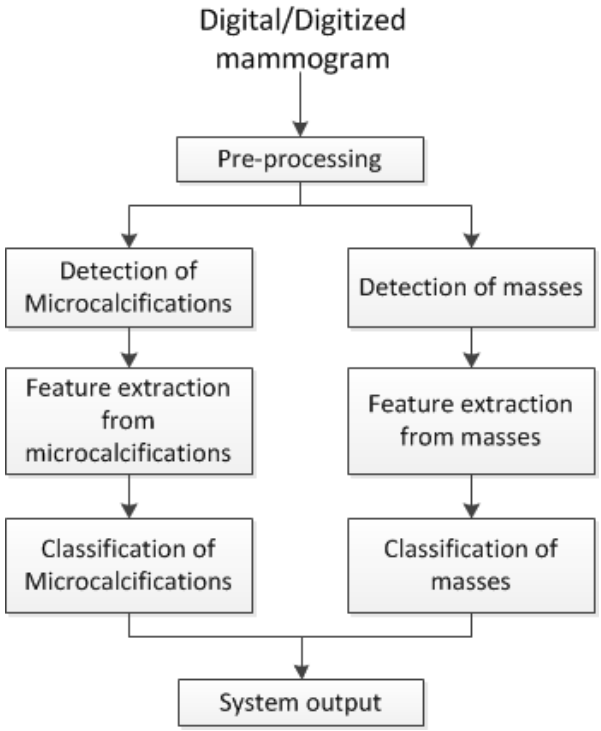


Figure 8.1: Experimentation methodology, designed under a linear, traditional CAD architecture.

set of 38 microcalcification clusters and 52 masses by performing the image segmentation techniques described in section 7.2.2. All experiments were performed in MATLAB Version 7.8.0.347 (R2009a), under a 2.8 GHz Intel Xeon processor with 3.48 GB of memory.

8.2 Feature Selection

In the case of microcalcification clusters, the evolutionary feature selection process that we designed (see Section 7.2.4), considered a GA that contained individuals with chromosomes of 31 bits of length, representing the inclusion (or exclusion) of each one of the 31 features extracted from the clusters. For mass-related features, individuals had a length of 50 bits.

In both cases, we used a GA with an embedded *diversity injection* mechanism, that considered the replacement of 90% of the current population by inserting new individuals with random chromosomes, if the evolutionary process did not find any enhancements in 10 consecutive generations.

Furthermore, it had a population of 100 individuals, binary tournament selection, two-point crossover and fitness based reinsertion. The probability of crossover was set to $p_c = 0.9$ for, both, cluster-features and mass-features; the probability of mutation was $p_m = 0.3$ in both processes. The initial population of the GA was initialized uniformly at random and ran for 200 generations.

This feature selection process was replicated separately for clusters and masses. In both executions, the process was performed three times, one for each of the classifiers of interest (k-NN, NN, SVM, LDA), taking the AUC as the fitness function for evaluating individuals.

The k-NN algorithm worked with a majority vote classification, considering a fixed $k = 11$ for both datasets. The NN had a single neuron in the output layer and complying with Kolmogorov's

theorem [59], one hidden layer with $2n + 1$ neurons, where n is the number of input units. All neurons had the sigmoid hyperbolic tangent as transfer function. As for the SVM, we used the Gaussian Radial Basis Function kernel, with a scaling factor of $\sigma = 2.0$ and the *Quadratic Programming* technique to find the separating hyperplane. Finally, we also considered the Linear Discriminant Analysis method to construct as the third classifier.

8.3 Classification Results for Clusters of Calcifications

Table 8.1 shows the subset of features obtained during the *wrapper-based selection* process for cluster-related features, considering the GA scheme described previously. This subset represents the best individual found in this evolutionary process, for each of the three classifiers of interest, independently. The k-NN and NN classifiers required 11 and 10 features, respectively, from all three categories listed in Table 7.1, being the largest subsets of features that were selected in this stage. On the other hand, both the SVM and LDA selected the same two features that describe the shape of the microcalcification cluster, namely: minimum radius and maximum diameter of the cluster.

Classifier	Selected Cluster Features
k-NN (11 features)	**Cluster shape:** number of calcifications, compactness, microcalcification density, mean of the distances between microcalcifications. **Area of Microcalcifications:** mean area of microcalcifications, standard deviation of the area of microcalcifications, relative area. **Microcalcification Contrast:** total gray mean level of microcalcifications, standard deviation of the mean gray levels of microcalcifications, total absolute contrast, standard deviation of the absolute contrast.
NN (10 features)	**Cluster shape:** compactness, microcalcification density, maximum radius, mean of the distances between microcalcifications. **Area of Microcalcifications:** Total area of microcalcifications, relative area. **Microcalcification Contrast:** Total gray mean level of microcalcifications, standard deviation of the mean gray levels of microcalcifications, total absolute contrast, minimum absolute contrast.
SVM (2 features)	**Cluster shape:** Minimum radius, maximum diameter.
LDA (2 features)	**Cluster shape:** Minimum radius, maximum diameter.

Table 8.1: Number of selected cluster features.

A leave-one-out cross-validation mechanism was embedded within the feature selection stage, with the objective of measuring the performance of the three classifiers regarding their ability to classify clusters of mircocalcifications. The overall accuracy, sensitivity and specificity were measured for each algorithm. Table 8.2 shows the results of the cross-validation, in which we can see that the highest overall accuracy (0.8974) was achieved by the NN, with a sensitivity of 0.6667 and a specificity of 0.9655, resulting in an $AUC = 0.8161$. On the other hand, both the SVM and LDA had the same performance on all three parameters, presenting a sensitivity of 1.000 and a specificity of 0.7241, which resulted in an overall accuracy of 0.7895 and $AUC = 0.8621$, which was the highest and represents the best results of this set of experiments.

Furthermore, we also present the amount of CPU time (in minutes) that each of the classifiers demanded, during this cross-validated feature-selection; the NN was the model which required the greatest amount of time, around 18.5 hours and the k-NN the one that demanded the less, terminating

in 10.2 minutes the whole selection and validation processes. Both the SVM and LDA required a much lower amount of time to complete than that of the NN, but the SVM demanded 5.8 hours to end its execution which is much greater than the LDA time and still provided the same classification performance.

Classifier	AUC	Accuracy	Sensitivity	Specificity	Time (min)
k-NN	0.7931	0.8158	0.7778	0.8276	10.2
NN	0.8161	0.8947	0.6667	0.9655	1110.54
SVM	0.8621	0.7895	1.000	0.7241	347.37
LDA	0.8621	0.7895	1.000	0.7241	17.62

Table 8.2: Cluster-classification Performance.

8.4 Classification Results for Masses

The same process performed in the feature selection and classification regarding microcalcification clusters was replicated for the detected masses, considering the aforementioned mass-related features and parameters of experimentation.

Table 8.3 presents the set of selected mass-related features, which represent the best individual found in the GA-based feature selection process, for each of the three classifiers of interest. In this case, all classification algorithms selected almost the same number of features; the k-NN used 7 features and the NN required 5 of them, while both the SVM and LDA used 4. We can observe that once again the k-NN required the largest subset.

Classifier	Selected Cluster Features
k-NN (7 features)	**Background contrast:** Minimum gray level, standard deviation of the gray level, gray level asymmetry (skewness). **Relative Contrast:** Absolute contrast. **Shape:** Area, background area, maximum diameter.
NN (5 features)	**Background contrast:** Standard deviation of the gray level, Gray level asymmetry (skewness). **Relative Contrast:** Absolute contrast. **Shape:** Area, background area.
SVM (4 features)	**Background contrast:** Gray level asymmetry (skewness). **Relative Contrast:** Absolute contrast, Proportional contrast. **Shape:** Background area.
LDA (4 features)	**Background contrast:** Standard deviation of the gray level. **Relative Contrast:** Proportional contrast. **Shape:** Convex area, background area.

Table 8.3: Number of selected masses features.

Finally, Table 8.4 shows the results of the leave-one-out cross-validation process, where it can be observed that the k-NN outperformed the other methods in terms of AUC. In this case The LDA presented the lowest performance, with a sensitivity of 0.6364 and a specificity of 0.6316, resulting in an overall mass-classification accuracy of 0.6346 and an $AUC = 0.6339$. On the other hand, the SVM was the second highest, providing an $AUC = 0.7169$, accuracy of 0.7115, a sensitivity of 0.6970

and a specificity 0.7368. The NN achieved the greatest sensitivity of all methods, but also the lowest specificity.

Classifier	AUC	Accuracy	Sensitivity	Specificity	Time (min)
k-NN	0.7624	0.7692	0.7368	0.7879	12.66
NN	0.6683	0.6923	0.5789	0.7576	1723.86
SVM	0.7169	0.7115	0.7368	0.6970	495.41
LDA	0.6339	0.6346	0.6316	0.6364	67.11

Table 8.4: Mass-classification Performance.

The amount of CPU time that each classifier needed to conduct this cross-validated feature-selection stage regarding masses, is also provided. We can see that once again the NN model required the greatest amount of time, of around 28.7 hours, while the SVM needed 8.25 hours to terminate its execution and the LDA 1.12 hours.

As we observed in this set of experiments, the results of the considered classifiers gives room for an enhancement, since the best performance in terms of AUC that was achieved with this methodology is 0.8621 and 0.7624, for MCCs and masses, respectively. Additionally, there are some classifiers that achieved a performance that is just slightly better than a random classifier, as it is the case of the NN in the masses dataset which reached 0.5789 sensitivity. On the other hand, in the MCCs dataset the considered algorithms either achieved a high sensitivity and a low specificity or the other way around, but they never achieved a high performance in both parameters. Therefore, in the next chapter we present the results that were obtained by implementing the proposed classification methodology, which has the primary objective of enabling classifiers to reach a better performance.

8.5 Summary

In this chapter we presented a first stage of experiments in which the objective was to measure the performance that can be achieved by the four classifiers of interest, under the traditional CAD architecture. For that, we designed and followed an experimentation methodology that performs all processes *sequentially*, following the conventional CAD approach.

We applied the computer vision techniques for lesion segmentation as well as the feature extraction stage and evolutionary feature selection as described in chapter 7. The classification stage was evaluated with a leave-one-out strategy and we present results for AUC, overall accuracy, sensitivity and specificity. The subset of selected features is also included in our results. We obtained the best results with the SVM and LDA for the classification of microcalcification clusters, reaching both methods an AUC of 0.8621, and the k-NN outperformed the rest of the classifiers in mass-classification, with an AUC of 0.7624.

Chapter 9

Classification Performance Applying the Proposed Model

This chapter presents the results that were obtained by applying the proposed solution model the way it was described in Chapter 7. The objective of performing these experiments is to show how competitive the proposed model is against a traditional CAD, which was tested in the previous chapter, and if implementing the CBR methodology that was proposed results in an enhancement of the classifiers' performance.

9.1 Experimentation Setup and Methodology

We implemented the same pre-processing and segmentation techniques that were described in section 7.2.2 and extracted from them the feature sets of Table 7.1 and Table 7.2. Based on the extracted features our system **retrieves** from the database a subset of k cases (or lesions) that are similar to the ones found in a query image, by performing a k-NN-based similarity search. Finally, the retrieved cases are **reused** to compute the diagnosis of the encountered lesions, based on the classifiers of interest.

This time the wrapper feature selection mechanism was executed with 200 individuals for 100 generations with a crossover probability of $p_c = 0.9$ and $p_m = 0.2$ mutation probability. The rest of the parameters remained the same: population initialized uniformly at random, binary tournament selection, two-point crossover, 31-bit long chromosomes for microcalcification clusters and 50 bits in every individual for tumoral masses. We also used the AUC as fitness function for individuals in the GAs.

These experiments have three objectives: (1) to determine the most suitable dissimilarity metric for the retrieval task, (2) to determine the optimal amount of k cases to be retrieved/re-used from the database and (3) to find the most relevant subset of features, in order to obtain the highest performance of each of the classification algorithms considered in this study. Therefore, we first explore the performance of six different dissimilarity metrics. Then, we perform feature selection and classification accuracy assessment for $k = 3, 5, 7, 9, 11, 13, 15, 17, 19, 21$. We then analyse which k value provides the highest performance and present the related subset of selected features.

The whole process was performed four times for every k, one for each of the classifiers of interest (k-NN, NN, SVM, LDA). The k-NN performed a majority-rule classification with respect to the k retrieved cases. The NN had one hidden layer with $2n + 1$ neurons, where n is the number of input units, according Kolmogorov's theorem [59], as well a single neuron in the output layer; all neurons

considered the sigmoid hyperbolic tangent as transfer function. Regarding the SVM, we used a scaling factor of $\sigma = 2.0$ in a Gaussian Radial Basis Function kernel and the *Quadratic Programming* technique to find the separating hyperplane. We also used the LDA method as a fourth classifier.

9.2 Evaluating Dissimilarity Metrics

In order to determine the most suitable dissimilarity metric to be used for the retrieval of similar cases from the database, we evaluated six different metrics in terms of the AUC that is obtained by performing a k-NN classification with different amounts k of retrieved instances.

Considering a query row vector $\mathbf{x}$ and any row vector $\mathbf{y}$ stored in the database, the dissimilarity metrics evaluated are:

- Euclidean distance:

$$d = \sqrt{(\mathbf{x}-\mathbf{y})\,(\mathbf{x}-\mathbf{y})'} \tag{9.1}$$

- Manhattan distance:

$$d = \sum_{i=1}^{n} |x_i - y_i| \tag{9.2}$$

- Linear correlation:

$$d = 1 - \frac{(\mathbf{x}-\bar{\mathbf{x}})(\mathbf{y}-\bar{\mathbf{y}})'}{\sqrt{(\mathbf{x}-\bar{\mathbf{x}})(\mathbf{x}-\bar{\mathbf{x}})'}\sqrt{(\mathbf{y}-\bar{\mathbf{y}})(\mathbf{y}-\bar{\mathbf{y}})'}} \tag{9.3}$$

- Cosine similarity:

$$d = 1 - \frac{\mathbf{x}\mathbf{y}'}{\sqrt{(\mathbf{x}\mathbf{x}')(\mathbf{y}\mathbf{y}')}} \tag{9.4}$$

- Mahalanobis distance:

$$d = \sqrt{(\mathbf{x}-\mathbf{y})\,\mathbf{C}^{-1}\,(\mathbf{x}-\mathbf{y})'} \tag{9.5}$$

  where $\mathbf{C}$ is the covariance matrix between $\mathbf{x}$ and $\mathbf{y}$.

- Spearman distance:

$$d = \frac{\sum_i (x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{9.6}$$

  where scores $\mathbf{x_i}$ and $\mathbf{y_i}$ are converted to ranks x_i, y_i.

Table 9.1 shows the results for this process, regarding the classification of MCCs. Each column shows the results of the AUC that was obtained by applying each of the considered metrics for k-NN classification and the highest AUC is underlined for each k.

We can see that the *correlation* metric consistently provided the highest accuracy, regardless of the amount of retrieved cases. It was only outperformed by the *spearman* metric ($k = 13$). They both

Classifier	k									
	3	**5**	**7**	**9**	**11**	**13**	**15**	**17**	**19**	**21**
Euclidean	0.8103	0.8103	0.7931	0.7759	0.7759	0.7069	0.6724	0.6207	0.5517	0.5345
Manhattan	0.8448	0.7931	0.7931	0.7759	0.7414	0.7069	0.6724	0.6379	0.5690	0.5172
Correlation	0.8621	0.8621	0.8103	0.7759	0.7931	0.7414	0.7759	0.7375	0.6341	0.5690
Cosine	0.8238	0.7720	0.7414	0.7586	0.7241	0.6897	0.6897	0.6724	0.5517	0.5345
Mahalanobis	0.8103	0.7931	0.7241	0.7241	0.6379	0.5862	0.5517	0.5172	0.5172	0.5172
Spearman	0.8027	0.8276	0.7893	0.7759	0.7893	0.8276	0.7203	0.6686	0.5862	0.5690

Table 9.1: Evaluation of dissimilarity metrics considering AUC obtained by k-NN classification of MCCs.

Classifier	k									
	3	**5**	**7**	**9**	**11**	**13**	**15**	**17**	**19**	**21**
Euclidean	0.7624	0.7209	0.7289	0.6986	0.6834	0.6946	0.7057	0.7209	0.6723	0.6723
Manhattan	0.7775	0.7440	0.7137	0.7209	0.7026	0.7209	0.6946	0.6946	0.6499	0.6986
Correlation	0.8381	0.7967	0.8038	0.7329	0.7624	0.7624	0.7624	0.7624	0.7663	0.7624
Cosine	0.8118	0.7967	0.7887	0.7361	0.7472	0.7624	0.7624	0.7472	0.7472	0.7361
Mahalanobis	0.6938	0.6252	0.5798	0.5606	0.5303	0.5152	0.5000	0.5000	0.5000	0.5000
Spearman	0.6611	0.6388	0.6459	0.6459	0.6388	0.6459	0.6459	0.5598	0.5303	0.5152

Table 9.2: Evaluation of dissimilarity metrics considering AUC obtained by k-NN classification of Masses.

achieved the same AUC for $k = 21$. For $k = 9$, all metrics achieved the same performance, except for *Mahalanobis* metric.

Table 9.2 shows the performance for the k-NN classification of masses for all the dissimilarity metrics considered. Once again, the correlation metric provided the highest AUC in almost all the experiments, except for $k = 9$, which *cosine* metric outperformed it. Moreover, for $k = 5, 13, 15$ both metrics achieved the same performance.

The dispersion of the AUC for MCCs and masses is depicted in Figure 9.1 and Figure 9.2, respectively. It can be observed that each metric has a different median performance and in both datasets correlation metric which presents the highest median.

We used Friedman's rank sum test [45] to determine if these performance differences are significant. Table 9.3 shows the mean ranks of AUC for each metric and the p-values of the test for both MCCs and masses. Both p-values (4.6e-09 and 4.49e-06) show significance at a 5% level.

Multiple comparisons tests as described in [84] were performed to verify the overall mean rank outperformance of the correlation metric. We compare the absolute mean rank difference between the correlation metric and the rest, $\left|\bar{R}_c - \bar{R}_i\right|$, against $z_{\alpha/(2k)}\sqrt{\frac{t(t+1)}{6n}} = 1.821$, where $z_{\alpha/(2k)} = 2.576$ is a standard normal quantile, with $\alpha = .05$, $k = 5$ comparisons, $t = 6$ number of treatments and $n = 10$ number of groups, since we are doing 5 comparisons among six metrics on ten blocks.

As shown in Table 9.4 and Table 9.5, the difference of performance between the correlation metric and the rest was significant in all cases, except for the cases in which it is compared to spearman (MCCs) and cosine metric (masses). All differences were positive. We conclude that the correlation metric has an overall tendency that outperforms the other metrics.

Based on these results, we determined to use the *correlation* metric in the similarity search process within the proposed CBR model, in which we train different classification algorithms with the set

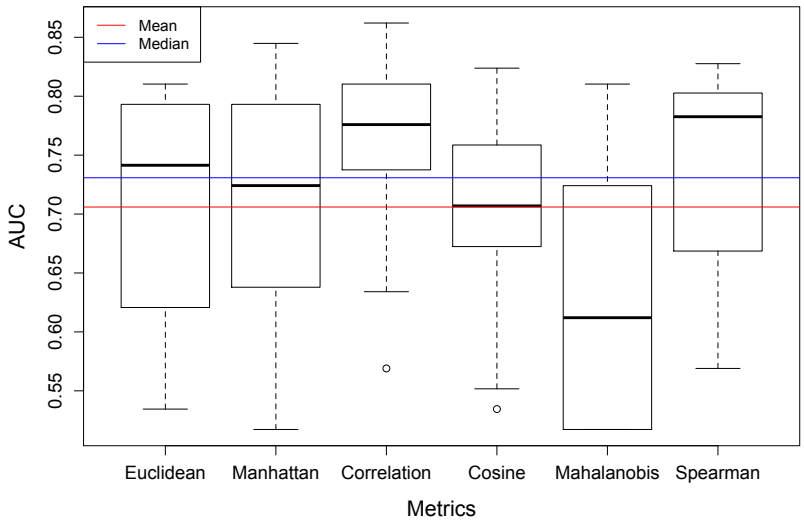


Figure 9.1: Boxplots of AUC performance for MCCs.

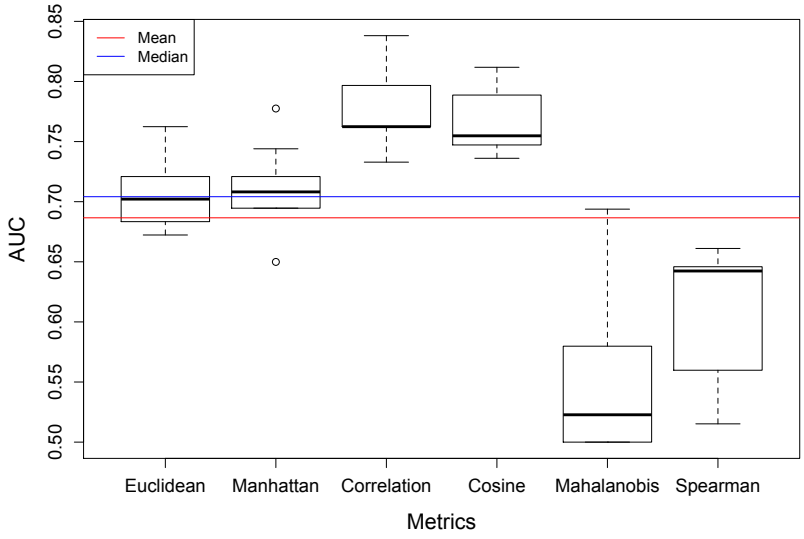


Figure 9.2: Boxplots of AUC performance for Masses.

of similar cases that are retrieved from the database of historical cases.

9.3 Classification Results for Microcalcification Clusters

We implemented a leave-one-out validation in which all the classifier algorithms were trained using the set of k similar cases retrieved from the database using the correlation metric. Table 9.6 shows the cross-validation results of ten different tests in which the four classifiers of interest were applied

Dissimilarity metric	Mean rank for MCCs	Mean rank for masses
Euclidean	3.35	3.40
Manhattan	3.40	3.60
Correlation	5.70	5.75
Cosine	2.80	5.25
Mahalanobis	1.35	1.10
Spearman	4.40	1.90
	p-value = 4.6e-09	p-value = 4.49e-06

Table 9.3: Results of Friedman Test.

Comparison	$\left\|\bar{R}_c - \bar{R}_i\right\|$
Correlation - Euclidean	2.35*
Correlation - Manhattan	2.30*
Correlation - Cosine	2.90*
Correlation - Mahalanobis	4.35*
Correlation - Spearman	1.30

* Significant at a familywise type I error rate of 5%.

Table 9.4: Pairwise comparison of metrics' mean ranks, with MCCs dataset.

Comparison	$\left\|\bar{R}_c - \bar{R}_i\right\|$
Correlation - Euclidean	2.35*
Correlation - Manhattan	2.15*
Correlation - Cosine	0.50
Correlation - Mahalanobis	4.65*
Correlation - Spearman	3.85*

* Significant at a familywise type I error rate of 5%.

Table 9.5: Pairwise comparison of metrics' mean ranks, with masses dataset.

to assess the malignancy of microcalcification clusters, training with different amounts of k retrieved samples, ranging from $k = 3$ to $k = 21$. We measured the AUC, overall accuracy, sensitivity and specificity, and the best results for each classifier, in terms of AUC, are underlined.

Figure 9.3 shows the dispersion of all four performance parameters for every classifier. We can see that the performance of the CBR model varies considerably among and within classificators. Every performance measure exhibits the presence of oultiers, specially the AUC. The NN and SVM classifiers have a median performance at or above the overall median on every measure considered, with one exception in the SVM's AUC, which is just slightly below. Also, these two classifiers tend to have a smaller dispersion. As a general conclusion, we may say that the performance of the proposed CBR model is sensitive to the classifier employed.

We can observe that the k-NN presented its best performance with $k = 3, 5$ with an $AUC =$

Classifier	Performance	k									
		3	**5**	**7**	**9**	**11**	**13**	**15**	**17**	**19**	**21**
k-NN	**AUC**	0.8621	0.8621	0.8103	0.7759	0.7931	0.7414	0.7759	0.7375	0.6341	0.5690
	Accuracy	0.7895	0.7895	0.7632	0.7632	0.8158	0.7368	0.7632	0.7895	0.6842	0.5789
	Sensitivity	1.0000	1.0000	0.8889	0.7778	0.7778	0.7778	0.7778	0.6667	0.5556	0.5556
	Specificity	0.7241	0.7241	0.7241	0.7586	0.8276	0.7241	0.7586	0.8276	0.7241	0.5862
NN	**AUC**	0.8161	0.8544	0.8927	0.9272	0.9100	0.9100	0.8889	0.8889	0.9100	0.9655
	Accuracy	0.8947	0.8947	0.8947	0.9474	0.9211	0.9211	0.9474	0.9474	0.9211	0.9474
	Sensitivity	0.6667	0.7778	0.8889	0.8889	0.8889	0.8889	0.7778	0.7778	0.8889	1.0000
	Specificity	0.9656	0.9310	0.8966	0.9655	0.9310	0.9310	1.0000	1.0000	0.9310	0.9310
SVM	**AUC**	0.8544	0.8333	0.8161	0.8544	0.8717	0.8717	0.8544	0.8372	0.8717	0.8544
	Accuracy	0.8947	0.9211	0.8947	0.8947	0.9211	0.9211	0.8948	0.8684	0.9211	0.8947
	Sensitivity	0.7778	0.6667	0.6667	0.7778	0.7778	0.7778	0.7778	0.7778	0.7778	0.7778
	Specificity	0.9310	1.0000	0.9656	0.9310	0.9656	0.9656	0.9310	0.8966	0.9656	0.9310
LDA	**AUC**	0.7989	0.8372	0.8717	0.8717	0.9310	0.8966	0.8793	0.8793	0.8582	0.8582
	Accuracy	0.8684	0.8684	0.9211	0.9211	0.8948	0.8421	0.8158	0.8158	0.8421	0.8421
	Sensitivity	0.6667	0.7778	0.7778	0.7778	1.0000	1.0000	1.0000	1.0000	0.8889	0.8889
	Specificity	0.9310	0.8966	0.9656	0.9656	0.8621	0.7931	0.7586	0.7586	0.8276	0.8276

Table 9.6: Classification Performance for Microcalcification Clusters

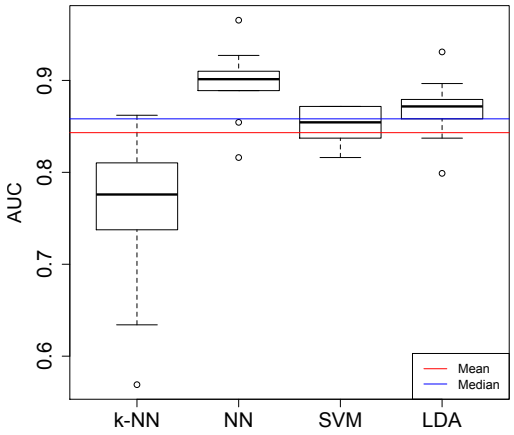
0.8621, which consistently decreased with higher amounts of retrieved cases k, as well as the sensitivity. Furthermore, we can see in Figure 9.3(a) that the sensitivity of the classifier seems to be highly unstable, since it presented the most variability. The performance of the rest of parameters was more stable. Is it worth mentioning the presence of outliers at $k = 21, 11, 17$, the first one with high influence on all tests and the latter on the specificity. Therefore, regarding this classifier, it can be observed that a more accurate and stable performance can be achieved considering a lower number of training samples, which is to be expected, given the simplistic approach of a majority-vote k-NN used in this experiment.

As for the NN, the best performance was obtained with $k = 21$, providing an $AUC = 0.9655$, an overall accuracy of 0.9474, 1.00 sensitivity and 0.9310 specificity. High results were obtained in terms of specificity and overall accuracy, with a low variability, as can be observed in Figure 9.3(b). On the other hand, the sensitivity of this classifier presented a high variability, indicating once again that the number of retrieved samples used for training the NN affects this parameter. Moreover, it achieved higher results in terms of AUC but presented outliers for $k = 3, 5, 21$.

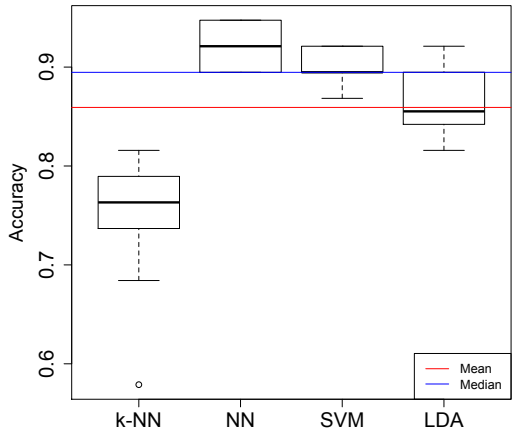
The SVM achieved its highest AUC, 0.8717%, with $k = 11, 13, 19$, resulting in an overall accuracy of 0.9211. Throughout the experiments the sensitivity achieved by this classifier was 0.7778 in all cases except for $k = 5, 7$, in which it decreased to 0.6667; however, these results are outliers, as can be observed in Figure 9.3(c). Moreover, it can be seen that the performance of this classifier was more stable across the different amounts of training samples k that were considered in the experiments.

As for LDA, the highest AUC was 0.9310, with 0.8948 overall accuracy, 1.000 sensitivity and 0.8621 specificity for $k = 11$. Moreover, the highest sensitivity was achieved for $k = 11, 13, 15, 17$, while the highest specificity, 0.9656, was obtained with $k = 7, 9$, in which also the highest overall accuracy of 0.9211 was achieved. We can observe in Figure 9.3(d) that this classifier's AUC presented outliers for $k = 3, 11$ with low variability, while both sensitivity and specificity were more disperse.

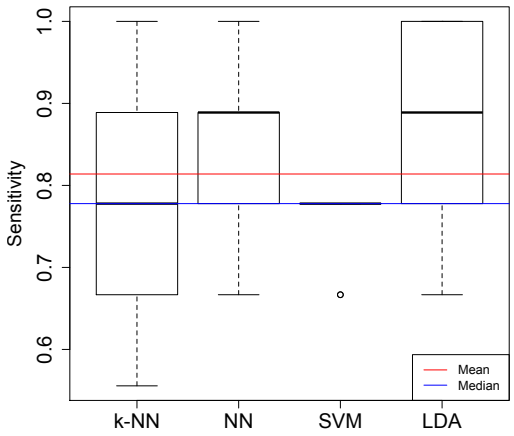
Table 9.7 presents, both, the subset of features and the size of k retrieved training samples with which we obtained the best classification performance regarding microcalcification clusters, based on the previous results. For k-NN we determined to use $k = 3$ which provides an AUC of 0.8621, since no enhancements were found with different values and, also, considering the retrieval and re-use of only 3 cases from the database represents less computational demand. Similarly, for the NN classifier we



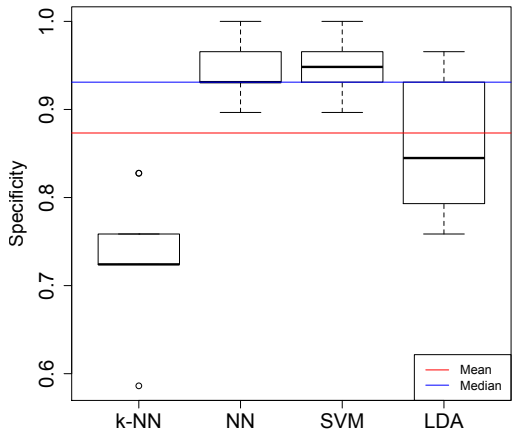
(a) Boxplots of AUC per classifier



(b) Boxplots of accuracy per classifier



(c) Boxplots of sensitivity per classifier



(d) Boxplots of specificity per classifier

Figure 9.3: Boxplots of performance measures for MCCs classification.

determined to use $k = 21$ with an AUC of 0.9655, 1.000 sensitivity and 0.9310 specificity, resulting in 0.9474 overall accuracy and for SVM and LDA we selected the results obtained with $k = 11$, with an AUC of 0.8717 and 0.9310, respectively.

9.4 Classification Results for Masses

The same GA-based feature selection process was applied for the diagnosis of masses. We performed the same leave-one-out cross-validation scheme that we described in the previous section. Table 9.8 presents the classification performance of the four classifiers considered, with respect to the diagnosis of tumoral masses; for each classifier, the best results in terms of AUC are underlined.

Boxplots in figure 9.4 show the dispersion of the various measures among classfiers. As with the

Classifier	k	Features Selected
k-NN	3	**Cluster shape:** convex perimeter, total radius, maximum radius, mean radius, standard deviation of radii, maximum diameter, mean of the distances between microcalcifications, standard deviation of the distances between microcalcifications. **Area of microcalfications:** total area, standard deviation of the area, relative area. **Microcalcification contrast:** total gray mean level of microcalcifications, standard deviation of the mean gray level of microcalcifications, mean absolute contrast, maximum absolute contrast.
NN	21	**Cluster shape:** convex area, microcalcification density, maximum radius, mean radius, minimum diameter.
SVM	11	**Cluster shape:** convex perimeter, convex area, minimum radius, minimum diameter.
LDA	11	**Cluster shape:** microcalcification density, total radius, minimum radius, minimum diameter.

Table 9.7: Features selected from Microcalcification Clusters

MCCs, it is apparent the great sensitivity of the model to the classifier used. In contrast to the MCCs, k-NN and LDA classifiers achieved a performance overall the median levels for each measure, except for the specificity achieved by k-NN, which is slightly below. We may note as well that the median performances of each measure differ from the MCCs results. The presence of outliers is also apparent from these plots.

Figure 9.4(a) shows that the k-NN classifier achieved its highest AUC performance, 0.8381, for $k = 3$, with 0.7368 sensitivity and a specificity of 0.9394 which was also the highest specificity of this classifier. We can also see that the variability of its performance was small, as opposed to the behavior that we observed in classifying microcalcification clusters. This is because the number of cases within this dataset is greater and the positive and negative classes have a low imbalance ratio.

The highest performance of the NN achieved an AUC of 0.7967, achieved with $k = 7$, resulting in 0.6842 sensitivity, 0.9091 specificity and an overall accuracy of 0.8269. In this case, the highest variability of this classifier's performance was observed in terms of sensitivity and preserved a high average specificity through all experiments, as can be seen in Figure 9.4(b).

As with the MCCs dataset, the SVM's performance remained stable in all runs of experimentation. This can be observed in Figure 9.4(c), which shows that the variability of this classifiers performance was relatively small in all metrics, except for its specificity. Additionally, the AUC, overall accuracy and specificity presented an outlier in $k = 21$, as well as its sensitivity in $k = 9$. Based on the AUC parameter, the best results of this classifier were obtained at $k = 5$, resulting in and AUC of 0.7815, overall accuracy of 0.8077, with a sensitivity of 0.6842 and a specificity of 0.8788

Finally, the highest performance of the LDA was achieved with $k = 7$, with an AUC of 0.8230, overall accuracy of 0.8462, a sensitivity of 0.7368 and 0.9091 specificity. Figure 9.4(d) shows that this classifier presented a high variability in terms of sensitivity and specificity, with outliers in overall accuracy and sensitivity, corresponding to $k = 7$ and $k = 11$, respectively.

Table 9.9 shows the set of selected features for the best run of each classifier (i.e. that in which the classifier achieved its best performance considering the AUC), which were selected based on the previous cross-validation results. We determined that the most desirable performance was obtained

Classifier	Performance	k									
		3	**5**	**7**	**9**	**11**	**13**	**15**	**17**	**19**	**21**
k-NN	**AUC**	0.8381	0.7967	0.8038	0.7329	0.7624	0.7624	0.7624	0.7624	0.7663	0.7624
	Accuracy	0.8654	0.8269	0.8077	0.7885	0.7692	0.7692	0.7692	0.7692	0.7885	0.7692
	Sensitivity	0.7368	0.6842	0.7895	0.5263	0.7368	0.7368	0.7368	0.7368	0.6842	0.7368
	Specificity	0.9394	0.9091	0.8182	0.9394	0.7879	0.7879	0.7879	0.7879	0.8485	0.7879
NN	**AUC**	0.6874	0.7289	0.7967	0.7855	0.7137	0.7289	0.7026	0.7097	0.6874	0.6834
	Accuracy	0.7308	0.7692	0.8269	0.8269	0.7500	0.7692	0.7500	0.7308	0.7308	0.7115
	Sensitivity	0.5263	0.5789	0.6842	0.6316	0.5789	0.5789	0.5263	0.6316	0.5263	0.5789
	Specificity	0.8485	0.8788	0.9091	0.9394	0.8485	0.8788	0.8788	0.7879	0.8485	0.7879
SVM	**AUC**	0.7624	0.7815	0.7472	0.7137	0.7209	0.7472	0.7281	0.7321	0.7169	0.6675
	Accuracy	0.7692	0.8077	0.7500	0.7500	0.7308	0.7500	0.7115	0.7308	0.7115	0.6346
	Sensitivity	0.7368	0.6842	0.7368	0.5789	0.6842	0.7368	0.7895	0.7368	0.7368	0.7895
	Specificity	0.7879	0.8788	0.7576	0.8485	0.7576	0.7576	0.6667	0.7273	0.6970	0.5455
LDA	**AUC**	0.8038	0.7512	0.8230	0.7815	0.8150	0.7815	0.7775	0.7663	0.7663	0.7624
	Accuracy	0.8077	0.7692	0.8462	0.8077	0.8077	0.8077	0.7885	0.7885	0.7885	0.7692
	Sensitivity	0.7895	0.6842	0.7368	0.6842	0.8421	0.6842	0.7368	0.6842	0.6842	0.7368
	Specificity	0.8182	0.8182	0.9091	0.8788	0.7879	0.8788	0.8182	0.8485	0.8485	0.7879

Table 9.8: Classification performance for masses

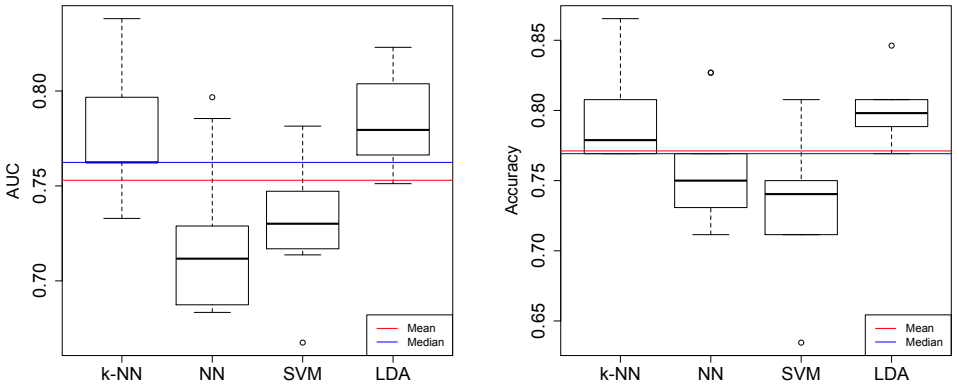
Classifier	k	Features Selected
k-NN	3	**Signal contrast:** Standard deviation of the gray level, Kurtosis of gray level. **Background contrast:** Maximum gray level, Minimum gray level.
NN	7	**Background contrast:** Median gray level, Mean gray level, Gray level asymmetry (skewness), Kurtosis of gray level. **Relative contrast:** 'Proportional contrast'. **Shape:** Area.
SVM	5	**Background contrast:** Minimum gray level, Mean gray level, Gray level asymmetry (skewness), Kurtosis of gray level.
LDA	7	**Background contrast:** Gray level asymmetry (skewness). **Relative contrast:** Absolute contrast, Relative contrast, Proportional contrast. **Shape:** Convex area, background area.

Table 9.9: Features selected from Masses

with $k = 3$ for the k-NN, resulting in a $AUC = 0.8381$, with $k = 7$ for NN and LDA algorithms presenting a 0.7967 and 0.8230 AUC, respectively and, finally, regarding the SVM, the optimal value was $k = 5$ with 0.7815 AUC.

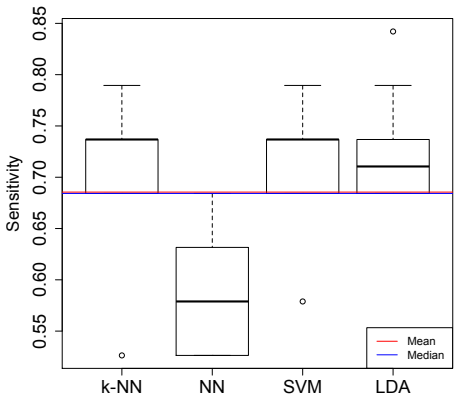
9.5 Comparison Against the Performance of the Traditional CAD

We have presented the results that can be reached by using within the proposed CBR methodology the same set of classifiers we used under the traditional CAD architecture. In this section we compare the performance of both approaches and discuss the enhancements we observe with the CBR model.

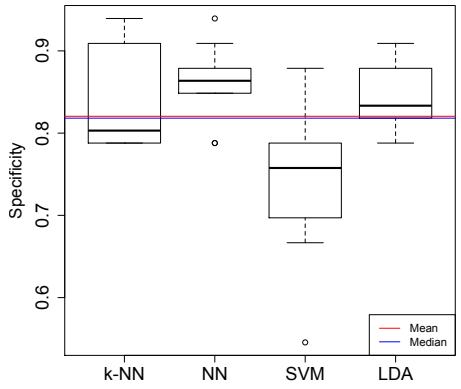


(a) Boxplots of AUC per classifier

(b) Boxplots of accuracy per classifier



(c) Boxplots of sensitivity per classifier



(d) Boxplots of specificity per classifier

Figure 9.4: Boxplots of performance measures for mass-classification.

Table 9.10 shows only the best performance observed for every classifier under the proposed classification methodology (underlined values in Table 9.6 and 9.8) and the results obtained by implementing the architecture of a traditional CAD, which were presented and discussed in chapter 8. The performance in terms of AUC is provided for both MCCs and masses.

In both datasets the CBR methodology designed in this research effort obtained the highest performance. In the case of MCCs classification, the **SVM** and **LDA** under the traditional scheme reached the same performance of the **k-NN** under our methodology ($AUC = 0.8621$), but that result represents the poorest one we obtained under the CBR model and, on the other hand, the highest of the traditional architecture. However, the rest of the classifiers within our scheme outperformed all of the algorithms used traditionally. The best result we observed out of all the considered classifiers was an outstanding $AUC = 0.9655$, achieved by the **NN** under CBR. Also, the LDA within our method reached a very high performance of $AUC = 0.9310$. On the other hand, the lowest AUC was achieved by the **k-NN** classifier in both approaches.

As for the masses dataset, the classifiers within our model outperformed their *traditional* counterparts. This time the **k-NN** reached the highest performance on both classification models, but within

CBR it still presented an enhancement, going from $AUC = 0.7624$ to $AUC = 0.8381$, in the traditional way and under CBR respectively. The *weakest* classifier under the proposed methodology was the **SVM** with $AUC = 0.7815$ and it still outperformed the *strongest* classifier in the traditional scheme which was the aforementioned **k-NN**.

Approach	Algorithm	AUC	
		MCCs	Masses
Traditional CAD	**k-NN**	0.7931	0.7624
	NN	0.8161	0.6683
	SVM	0.8621	0.7169
	LDA	0.8621	0.6339
CBR model	**k-NN**	0.8621	0.8381
	NN	0.9655	0.7967
	SVM	0.8717	0.7815
	LDA	0.9310	0.8230

Table 9.10: Comparison of performance between the proposed model and classifiers under the traditional CAD pipeline.

It should be noted that if we compare the results of every classifier individually, all of them increased their performance in both datasets under the CBR classification methodology that is proposed in this research work. The most notable enhancement was achieved by the **NN** in the MCCs dataset, going from $AUC = 0.8161$ to $AUC = 0.9655$ and by the **LDA** in the masses dataset, increasing its AUC from 0.6339 to 0.8230. The lowest performance enhancement was observed in **SVM** classifier in the MCCs dataset: from 0.8621 to 0.8717.

9.6 Visual Output of the Proposed Model

Figure 9.5 and 9.6 show the pictorical or visual output that the proposed model presents to the radiologist, for diagnosing MCC and masses respectively. Two representative cases from both lesions were selected to serve as queries to the system. The linear correlation, described in equation 9.3, was used to compute the similarity between the feature vectors of the query lesion and each of those stored in the database; afterwards, the system ranks all cases in the database sing this similarity score and, finally, takes the k most similar ones to train the classifiers, individually. For each classification algorithm and in descending order from left to right, we present in the mentioned figures the top three most similar cases for each representative lesion, i.e. those cases that were found to be most similar to the query case, considering the feature set that was selected by each classifier.

Therefore, it should be noted that the similar images that are displayed are not necessarily the same from one classifier to the other, since the similarity between the query and images in the database is calculated using the subset of features that each classifier selected and, consequently, the similarity search process is performed in different feature spaces. Refer to tables 9.7 and 9.9 to recall the features selected by each classification algorithm.

However, the image retrieval mechanism do *selects* the same images for different classifiers, as it is the case of the lesions depicted in Figure 9.5(c) and 9.5(l), as well as those illustrated in Figure 9.5(e) and 9.5(k), and also Figure 9.5(f) and 9.5(j), in the MCCs dataset.

(a) Sample MCC used as query.

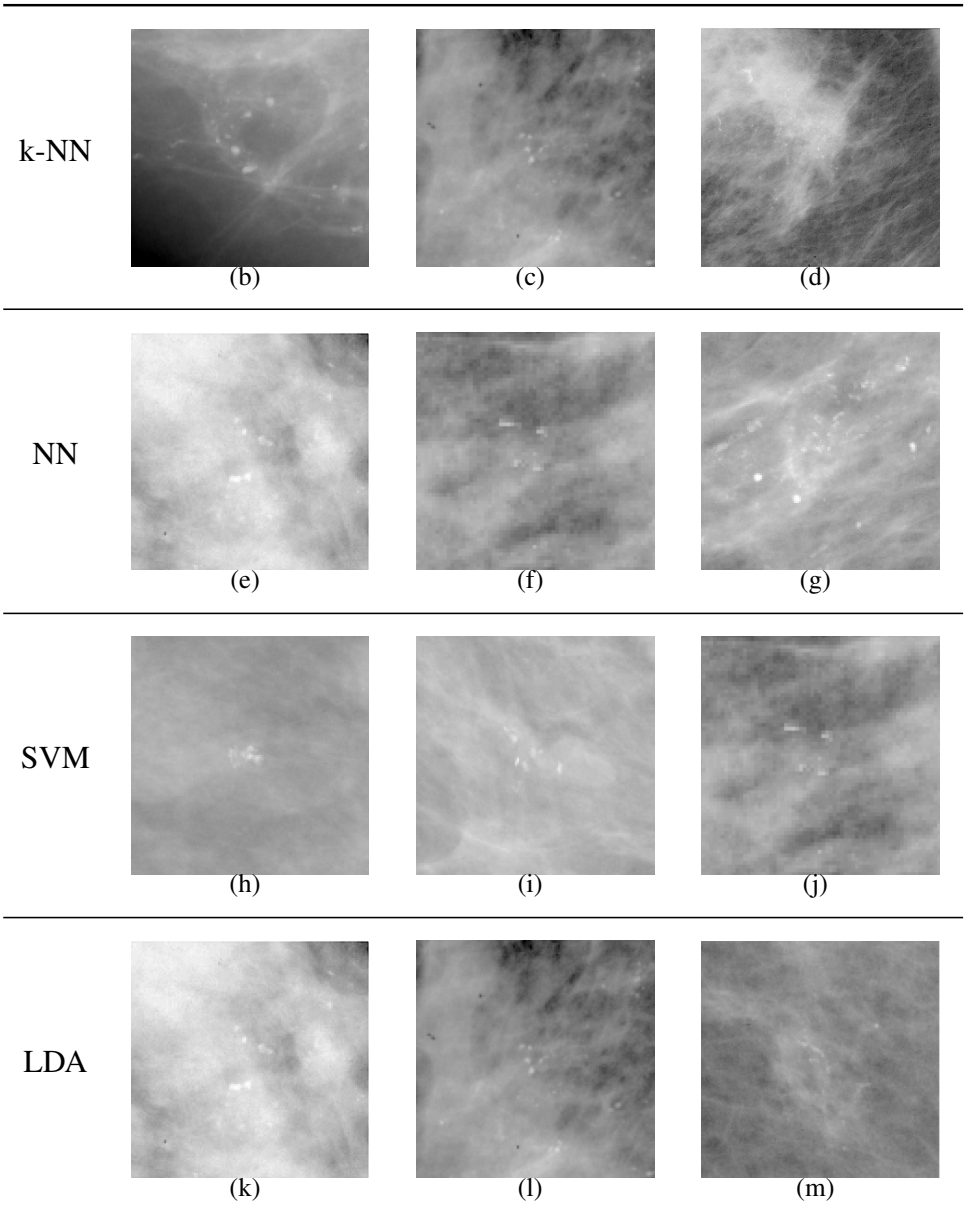
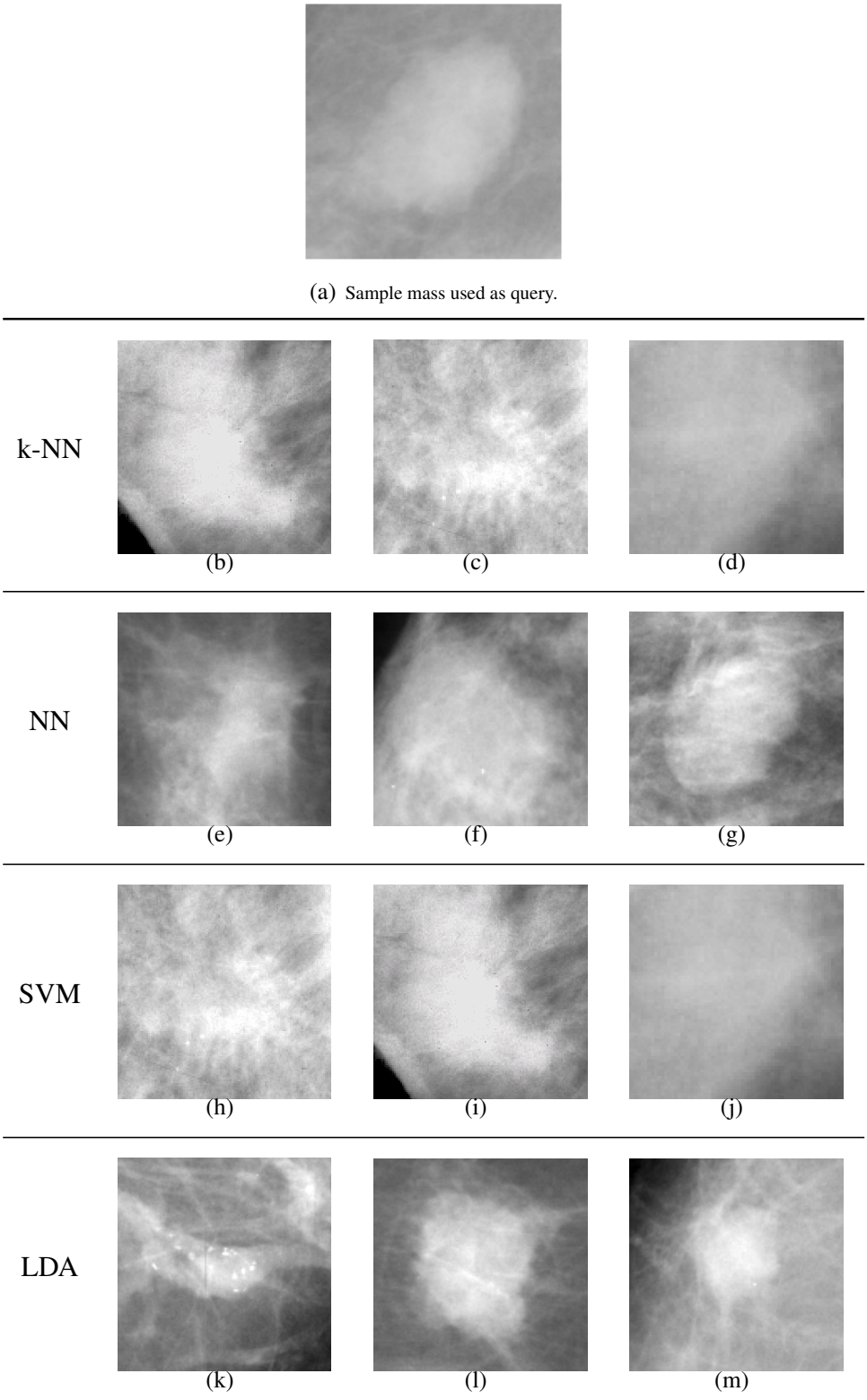


Figure 9.5: Example of the pictorical output for a representative MCC. The lesion in Figure 9.5(a) was used as query, corresponding to mdb219, from MIAS database. The top three most similar lesions are displayed for each classifier. Left-most images are more similar to the query than right most images, based on the *correlation* factor computed with features selected by each classifier (see Table 9.7).



(a) Sample mass used as query.

(b) (c) (d)

(e) (f) (g)

(h) (i) (j)

(k) (l) (m)

Figure 9.6: Example of the pictorical output for the representative mass shown in Figure 9.6(a), corresponding to image mdb015 from MIAS database. Top three most similar cases are displayed, from left to right in descending order. Similarity was computed by the correlation metric using features in Table 9.9 that were selected by each classifier.

Moreover, in the masses dataset, we can also observe that different classifiers used the same cases for training, in this particular example. Actually, the system retrieved the same set of images for the k-NN and SVM, although they were ranked in a different order since they belong to different feature spaces, as it was already mentioned. On the other hand, in both examples of the visual output, we can observe that the classifiers often use a diverse set of training cases, which introduces diversity to the system by exploring the feature space under different perspectives.

The objective of displaying these images to the end user is two-fold:

1. To serve as a visual or pictorical justification of the system's answer and prevent the resulting CAD from being a black box to physicians, who do not necessarily know the underlying processes of classification algorithms in order to understand their answers. With this visual output they will get to know the *knowledge base* that was used to compute the automatic diagnosis of the system and will be able to understand, at some extend, the system's suggested diagnosis.

2. To provide the radiologist with more information that can be taken into account in the diagnosis process, since the system not only provides the assessment of malignancy alone, but a set of historical cases with validated pathologies that are similar to the one in study and which can be regarded as pieces of information that, combined with the physician's expertise and the system's suggested diagnosis, will ultimately increase the chances for the radiologist to issue an accurate final diagnosis.

9.7 Summary

In this chapter we presented a stage of experiments in which the four classifiers of interest were trained under the CBR architecture that we previously described, in order to compare the performance that the classification algorithms can achieve with this strategy against the results that were obtained with the traditional approach in chapter 8.

A set of six similarity metrics were evaluated to determine the one that provides the most accurate results for our case-retrieval mechanism, considering the AUC that a simple k-NN can reach by using each metric; we used the Friedman's rank sum test to determine if the observed performance differences were significant and further evaluating with multiple multiple comparisons tests to confirm the overall mean rank outperformance of the correlation metric. It was found to be significant in all cases, except when compared against the spearman metric and the cosine similarity, for MCCs and masses, respectively.

Consequently, we determined to use correlation metric for the classification stage, in which we tested the four classifiers within the CBR methodology, implementing all phases that were described in the previous chapter 7. We determined with these experiments the optimal amount of instances that should be retrieved from the database to obtain the highest performance for each classifier, and presented the results in terms of AUC, overall accuracy, sensitivity and specificity.

In the MCCs dataset, the best results were obtained with $k = 3$ for k-NN and $k = 21$ for the NN; furthermore, both the SVM and LDA achieved their best results with $k = 11$. The performance results in terms of AUC that were obtained in this dataset were 0.8621, 0.9655 0.8717 and 0.9310, respectively. On the other hand, for mass classification the k-NN outperformed the rest of the classifiers reaching an AUC of 0.8381 with $k = 3$, while the NN, SVM and LDA achieved an AUC of 0.7967, 0.7815, 0.8230, with $k = 7$, $k = 5$ and $k = 7$, respectively.

Moreover, when we compared these results against those obtained with the traditional scheme, we concluded that implementing the classification of both breast cancer lesions under the proposed

approach resulted in a enhanced performance for all the tested classifiers, which describes a high potential of our model as a classification methodology and as a solution for breast cancer diagnosis.

Finally, we presented the visual output of the system, which has the objective of justifying the answer of the system in order to prevent our CAD from being a black box to physicians and, also, to provide them with information of cases that were studied in the past (which already have a validated pathology and contain similar lesions as their query) for them to consider in making a final diagnosis decision.

Chapter 10

Comparing the Proposed Model Against Classifier Ensembles and Imbalanced Learning Algorithms

In this chapter we will present the performance that we get by generating classifier ensembles and combining them with learning algorithms that are able to deal with the class-imbalanced feature of almost any medical dataset, which is related to the presence of a greater number of instances from (typically) the negative class and very few cases that belong to the positive one. In the particular problem that we are addressing, this means that our dataset has more negative breast cancer cases than positive ones and, ultimately, this could make the classifiers bias their answers towards the majority (negative) class, resulting in a low classification performance.

The objective of conducting these experiments is to be able to compare the proposed model against alternative classification approaches that can be used within the breast cancer diagnosis problem that we are addressing and introducing a discussion on how competitive the model is against related methods. Hence, in this chapter we will present the results we achieve with the aforementioned ensembles applied to our breast cancer datasets and then we will compare them against the performance of the CBR methodology, which was discussed in chapter 9.

10.1 Experimentation Setup

As it should be recalled, we are working with two datasets that contain cases from microcalcification clusters and breast masses. The former dataset contains 38 instances of that lesion, of which 9 are positive cases and 29 have a negative cancer diagnosis, resulting in an imbalance ratio of $IR = 3.222$. As for masses, there are 52 examples, 19 of them being positive cancer cases and 33 of them negative ones; hence, this dataset has an imbalance ratio of $IR = 1.736$. Table 10.1 presents a summary of this information.

Dataset	Total number of instances	Positive instances	Negative instances	Imbalance Ratio (IR)
Microcalcification clusters	38	9	29	3.222
Masses	52	19	33	1.736

Table 10.1: Imbalance ratio of datasets

Ensemble	Parameters
Bagging	FURIA as base classifier. A total of 10 classifiers in the ensemble.
Random Subspace	Sampling 25% of the feature space. FURIA as base classifier. A total of 10 classifiers in the ensemble.

Table 10.2: Parameters used to build the classifier ensembles.

In these experiments we built classifier ensembles considering 10 instances of the FURIA algorithm under Bagging, Random Subspace and also a combination of both approaches. These homogeneous classifiers were also combined with imbalance learning techniques such as SMOTE and the ENN rule. We set the Random Subspace algorithm to use a sampling rate of 25% of the feature set for each individual FURIA classifier in the ensemble. Table 10.2 summarizes the configuration we used for building these models.

Based on the previous combinations we evaluated a set of 6 classifier ensembles: Bagging and Random Subspace (individually), then we built the ensemble combining Bagging and Random Subspace and, finally, we introduced the SMOTE and ENN rule to the three ensemble approaches we are considering.

For every experiment, we present different performance metrics including AUC, accuracy, sensitivity and specificity, but we will strongly focus AUC in discussing these results, since its an important metric for experiments with imbalanced datasets. The rest of the metrics are included for completeness and coherence with experiments in the previous chapters; they will be taken into account in our discussions only to provide more information, when necessary, but not to make conclusions.

10.2 Classification Results for Microcalcification Clusters

Table 10.3 shows the classification results for the MCCs dataset. We can observe that using the Bagging and Random Subspace ensembles, without any combination with other methods, we achieve the lowest classification AUC: 0.7471 and 0.7854, respectively. However, results improve if we either combine both algorithms or introduce **SMOTE+ENN** to deal with the class-imabalanced issue, which means that both making structural changes in the ensemble (the way the are built) and processing the training data to balance both classes is key in determining the reaching a higher performance.

For the combination of Bagging and Random Subspace we obtain $AUC = 0.8199$, which is greater than both of the results we observe for those algorithms individually. On the other hand, applying Bagging and processing the dataset with SMOTE and the ENN rule (**Bagging** $_{\textbf{SMOTE+ENN}}$) results in the exact same performance as the one we get with the combination of both classifier ensemble techniques, but it is still higher than using Bagging alone.

Additionally, using **SMOTE+ENN** yields a higher performance for Random Subspace as well, resulting in $AUC = 0.8027$ as opposed to $AUC = 0.7854$ achieved with that ensemble approach alone. Finally, for the case in which we combine Bagging, Random Subspace and SMOTE+ENN (**B + RS + SMOTE + EN**), we obtained $AUC = 0.8409$ which is the highest performance we achieved for these set of experiments. Also, this ensemble yielded the highest sensitivity out of all the considered algorithms, which describes a high potential for correctly classifying positive cases.

Classifier	AUC	Accuracy	Sensitivity	Specificity
Bagging	0.7471	0.7895	0.6667	0.8276
Random Subspace	0.7854	0.7895	0.7778	0.7931
Bagging + Random Subspace	0.8199	0.8421	0.7778	0.8621
Bagging $_{\text{SMOTE+ENN}}$	0.8199	0.8421	0.7778	0.8621
Random Subspace $_{\text{SMOTE+ENN}}$	0.8027	0.8158	0.7778	0.8276
B + RS + SMOTE + EN	0.8409	0.8158	0.8889	0.7931

B = Bagging, RS = Random Subspace.

Table 10.3: Performance of learning algorithms for imbalanced datasets applied to MCCs dataset.

10.3 Classification Results for Masses

In Table 10.4 we present the performance of the same six algorithms, applied to the masses dataset. Again we can observe that using Bagging and Random Subspace for building the classifier ensemble without including any other of the considered techniques, provides the lowest results; in fact, they are just slightly better than the AUC that would be expected from a random classifier: 0.5510 and 0.5598, respectively.

Furthermore, the ensemble that was built with the combination of Bagging and Random Subspace achieves an $AUC = 0.5667$, which is better than the performance reached with the individual algorithms, but still very low and slightly higher than a random classifier as well. Also, the enhancement in performance is not that significant, since the AUC increases in roughly 1%.

On the other hand, the inclusion of SMOTE and ENN for the ensemble that was built under the Bagging approach (**Bagging** $_{\text{SMOTE+ENN}}$) provides an $AUC = 0.6012$ which represents a higher performance than its individual counterpart, and the same happens with the combination of Random Subspace with SMOTE and ENN (**Random Subspace** $_{\text{SMOTE+ENN}}$). However, it can be observed that for the latter ensemble the enhancement of performance that results from introducing SMOTE and ENN is of roughly 2%, resulting in an $AUC = 0.5774$ which is, again, no that much higher than a random classifier.

The highest performance was once again achieved by implementing a combination of Bagging and Random Subspace for building the classifier ensemble, and also using SMOTE and ENN for processing the training data (**B + RS + SMOTE + EN**). Its AUC, 0.7544, is way higher than the rest of the algorithms and, also, this ensemble clearly outperforms the rest of the approaches with 0.8421 sensitivity and 0.6667 specificity.

10.4 Comparison Against the Performance of the Proposed Model

Table 10.5 presents the results of the proposed CBR-based model and those related to the six classifier ensembles that were previously described. It shows the results that were achieved in both datasets (MCCs and masses) by all classifiers that were considered for the CBR methodology and we only focus in comparing the obtained AUC, since it is the most important metric for class-imbalanced learning

Classifier	AUC	Accuracy	Sensitivity	Specificity
Bagging	0.5510	0.5577	0.5263	0.5758
Random Subspace	0.5598	0.5385	0.5263	0.5455
Bagging + Random Subspace	0.5667	0.5800	0.5000	0.6333
Bagging $_{\text{SMOTE+ENN}}$	0.6012	0.5782	0.6500	0.5524
Random Subspace $_{\text{SMOTE+ENN}}$	0.5774	0.5769	0.5789	0.5758
B + RS + SMOTE + EN	0.7544	0.7308	0.8421	0.6667

B = Bagging, RS = Random Subspace.

Table 10.4: Performance of learning algorithms for imbalanced datasets applied to masses dataset.

algorithms and also represents a robust metric for comparing different machine learning algorithms. Only the best results were included for the algorithms used under the CBR model; refer to Table 9.6 and 9.8 to recall the complete set of results for MCCs and masses, respectively.

It can be observed that the algorithms that were used within the proposed CBR methodology achieved the highest performance in terms of AUC, for both datasets. Considering the results for the MCCs dataset, the *weakest* algorithm under the CBR model was the k-NN and with an $AUC = 0.8621$ still outperforms the combination of Bagging, Random Subspace, SMOTE and ENN (**B + RS + SMOTE + EN**), which was the *strongest* algorithm of the ensembles with $AUC = 0.8409$. As for the masses dataset, this ensemble was once again the more accurate but performed poorer than the weakest algorithm of the CBR methodology which was the **SVM** classifier with $AUC = 0.7815$.

On the other hand, the highest performance was achieved by the **NN** classifier in the MCCs dataset, achieving an $AUC = 0.9655$, and by the **k-NN** in the masses dataset with $AUC = 0.8381$, both of which were used under CBR. Moreover, the performance of the algorithms that were used within the proposed methodology shows a lower variability than the classifier ensembles.

Approach	Algorithm	AUC	
		MCCs	Masses
CBR model	**k-NN**	0.8621	0.8381
	NN	0.9655	0.7967
	SVM	0.8717	0.7815
	LDA	0.9310	0.8230
Ensembles	**Bagging**	0.7471	0.5510
	Random Subspace	0.7854	0.5598
	Bagging + Random Subspace	0.8199	0.5667
	Bagging $_{\text{SMOTE+ENN}}$	0.8199	0.6012
	Random Subspace $_{\text{SMOTE+ENN}}$	0.8027	0.5774
	B + RS + SMOTE + EN	0.8409	0.7544

Table 10.5: Comparison of performance between the proposed model and classifier ensembles with imbalanced learning techniques.

However, it is important to mention that the experiments were conducted with datasets that are limited in the number of instances and, hence, there is a possibility that the advantages of using classifier ensembles and introducing techniques to deal with the class imbalance issue might not have been fully exploited. Therefore, further experimentation should be conducted using larger datasets to make conclusive assertions about how competitive the proposed CBR model is against related approaches. Nonetheless, we can observe with the experiments presented in this section that introducing imbalance learning algorithms does result in a higher performance, which encourages the fact to keep exploring these techniques in further experiments.

10.5 Summary

In this chapter we presented a third stage of experiments which have the objective of testing machine learning methods that can serve as an alternative solution for breast-cancer diagnosis. We built a set of six different classifier ensembles using Bagging and Random Subspace, by combining them with each other and also introducing SMOTE+ENN as techniques to deal with the class-imbalance problem.

We found that the highest performance was achieved with the combination of Bagging, Random Subspace and SMOTE+ENN, in both datasets. Also, we observed that the ensembles that used SMOTE+ENN presented an enhanced performance in all cases, as compared to the version of the ensemble that did not use these resampling techniques, which implies that these algorithms benefited from dealing with the class-imbalance problem.

However, the classifiers that we used within the proposed model outperform all of the classifier ensembles that were tested in this chapter, but, it is important to clarify that number of examples in the datasets that we used for these experiments, do not allow to make conclusive statements on this matter, as the classifier ensembles and techniques for learning from imbalanced datasets have proved to be effective in several domains and further experiments with larger datasets should be conducted to make final conclusions. We should limit to state that, in this particular domain (with our datasets), the experiments demonstrate that our model presented a much higher performance.

Chapter 11

Conclusions and Future Work

This chapter will introduce the conclusions that can be drawn from the research effort described in this Dissertation. We proposed a classification methodology that was applied to the detection and automatic diagnosis of two important breast lesions: microcalcification clusters (MCCs) and masses. These anomalies can be evidence of malignancy in early breast cancer stages and, therefore, it is mandatory to perform clinical exams using screening technologies that include mammography, ultrasound, MRIs and many others. However, mammography is typically the first step in the detection process since it reveals important breast features to the radiologist that are useful to diagnose breast cancer in the earliest stages, when it can still be cured.

The proposed model is inspired in the Case-Based Reasoning (CBR) philosophy, which works under the premise that similar problems have similar solutions. The output of a CBR system is computed based on the information that can be re-used from historical cases that are stored in a knowledge-base and which have similar features to the query case; such cases were previously examined by an expert and contain solutions that are regarded as correct and validated and, therefore, they are retrieved from the database and processed, or re-used, in an appropriate way to take advantage from their solutions and come up with an answer to the new problem. Then, the new case along with its solution is retained in the database for future use, after being revised and validated by an expert.

In our experimental study we proposed an automatic way to segment MCCs and masses using computer vision and image processing techniques that were applied to mammographies from the MIAS database, which is widely used in research projects related to computer-aided breast cancer detection and diagnosis. Additionally, we successfully implemented two key processes of the CBR methodology: *retrieval* and *re-use* of similar cases that are stored in a database of historical data and applied them to the classification of MCCs and masses that were previously encountered by our lesion segmentation procedures.

We explored the performance of six different dissimilarity metrics applied to compute the retrieval of similar cases within an indexed *k-nearest-neighbors* similarity search, based on the vectors of visual features extracted from the lesions related to MCCs and masses. We used the AUC of a k-NN classifier to evaluate the precision of dissimilarity metrics and, afterwards, we used the Friedman test to determine if the difference of performance between all of them was significant. We then performed a pairwise comparison of the metrics' mean ranks to test if the overall mean outperformance of the *correlation* metric was statistically significant. Based on these results, we determined to use *correlation* for our similarity search for both datasets.

We tested our CBR architecture with four different types of classifiers, in order to explore the performance that the model could obtain with different machine learning perspectives: (1) the NN, which is a *non-linear* classifier that has been used in several classification domains and proved to be

highly adaptable, (1) the SVM, on the other hand, was selected to include a *kernel-based* classifier that has strong theoretical foundations, a good generalization capability and has also become popular given the high-performance it has achieved in different domains; (3) the LDA was included to explore a *linear* scheme and (4) the k-NN represents a simplistic approach based on a *majority-vote* classification mechanism.

A leave-one-out cross-validation was carried on the four considered classifiers algorithms (k-NN, NN, SVM and LDA), which were trained by *re-using* a set of k similar cases retrieved from our historical database. We considered several runs for this test in order to explore the performance of the classifiers across training sets of different sizes to finally be able to determine the amount of training samples that provided the best performance for each algorithm. Regarding MCCs, the best performance was obtained with $k = 5$ for k-NN, $k = 21$ for NN and $k = 11$ for SVM and LDA, with an AUC of 0.8621, 0.9655, 0.8717 and 0.9310, respectively. On the other hand, the NN and LDA with $k = 7$ presented their highest mass-classification performance, while the k-NN performed best with $k = 3$ and the SVM with $k = 5$, achieving an AUC of 0.7967, 0.8230, 0.8381 and 0.7815, respectively.

We compared our results against the performance that can be achieved with the traditional CAD pipeline and observed that introducing our CBR methodology results in an increase of performance for all of the four considered classifiers. Therefore, we can conclude that our framework fits best as a solution for accurately classifying breast cancer lesions. Additionally, we presented a set of experiments in which we compared six classifier-ensemble techniques that we built using Bagging and Random Subspace, combining each other and also introducing SMOTE and ENN to pre-process the training data to deal with the class-imbalanced issue and proved that our CBR classification methodology provides better classification results.

Table 11.1 shows a summary of all the results that were obtained from the three experimentation phases that were conducted in this research effort, in which we tested (1) the performance that could be obtained by the selected classifiers under a traditional CAD architecture, (2) under the proposed CBR methodology and, also, (3) the performance of alternative machine learning mechanisms such as classifier ensembles.

It can be observed that the CBR model achieved the best results in both datasets, showing a high performance that could not be achieved by any classifier of the other two approaches. The sole exception is the SVM and LDA of the traditional scheme in the MCCs dataset, which reached the same performance as the k-NN under the CBR model, but even though it is the best result obtained by the traditional architecture, it represents the lowest performance we achieved with the proposed solution. As for the masses dataset, the classifiers under the CBR methodology outperformed every other classifier of the other approaches.

It is important to note that all four classifiers in the traditional approach increased their performance when we used them within our methodology, which is evidence to the fact that the solution model holds a high potential of enabling a performance-enhancement in the algorithms that are used within the model. A very important enhancement was observed in the NN used in MCCs classification, which increased its AUC from 0.8161 to 0.9655. Also, in the masses dataset the LDA went from an AUC of 0.6339 to 0.8230.

The best results of the tested classifier ensembles were achieved by the combination of Bagging, Random Subspace, SMOTE and the ENN rule. This classifier reached an AUC of 0.8409 in the MCCs dataset, which can be considered a good performance on its own. In fact, it outperformed the k-NN and NN of the traditional approach and was close to reaching the best performance (0.8621) of that set of traditional classifiers. Moreover, in the masses dataset, this ensemble achieved an AUC of 0.7544, outperforming the NN, SVM and LDA classifiers of the traditional scheme and, once again, staying close to reaching the best classifier of that approach, which was the k-NN with an AUC of 0.7624. It is also important to mention that introducing SMOTE and ENN resulted in higher performance for all of

Approach	Algorithm	AUC	
		MCCs	Masses
Traditional CAD	**k-NN**	0.7931	0.7624
	NN	0.8161	0.6683
	SVM	0.8621	0.7169
	LDA	0.8621	0.6339
CBR model	**k-NN**	0.8621	0.8381
	NN	0.9655	0.7967
	SVM	0.8717	0.7815
	LDA	0.9310	0.8230
Ensembles	**Bagging**	0.7471	0.5510
	Random Subspace	0.7854	0.5598
	Bagging + Random Subspace	0.8199	0.5667
	Bagging $_{\text{SMOTE+ENN}}$	0.8199	0.6012
	Random Subspace $_{\text{SMOTE+ENN}}$	0.8027	0.5774
	B + RS + SMOTE + EN	0.8409	0.7544

Table 11.1: Summary of results obtained in the three different sets of experiments.

the ensembles that were tested and, consequently, encourages to further explore imbalanced learning techniques.

On the other hand, the number of instances of our datasets should be increased in order to study further the potential ability of enhancing the performance of integrated classifiers that is currently observed in our CBR model. Also, with larger datasets we could be able to conduct conclusive experiments on how competitive our model is against related approaches, such as the classifier ensembles that were tested in this study. It will also enable to fully exploit the advantages of imbalance learning techniques.

Finally, the proposed solution model not only provides the assessment of malignancy to the radiologist, but also displays the set of historical similar cases that were retrieved from the database. The objective of doing this is, on one hand, to show a visual *justification* of the system's output and, hence, preventing our CAD from being a black box to clinicians who are not necessarily familiar with the technical aspects of the inner phases of this type of systems and, therefore, do not always understand their answers. We claim that physicians will profit from this pictorical answer, since they will see the knowledge base upon which the classification of cases was conducted and this will enable them to understand the system's output; also, knowing what information was used to compute the suggested diagnosis gives physicians the choice to decide being confident or not with the system's answer.

On the other hand, we claim that this process provides the radiologist with more information that can be useful in the diagnosis process, since they can take into consideration not only the suggested malignancy, but also several pieces of historical, validated information that are similar to the case in study. Combining their expertise with both of the system's output will increase the chances of making an accurate diagnosis.

11.1 Future Work

The classification methodology that was designed in this doctoral dissertation proved to be able to classify breast cancer lesions with high accuracy and it has opened several research avenues that can be explored in future projects, including:

1. To explore alternative computer vision and image processing techniques, that can be used to segment lesions from digital mammographies and present the obtained results to an expert radiologist to provide a visual evaluation.

2. To extend the number of cases by using a larger database, in order to conduct experiments from which to draw strong conclusions about the models that will be tested. It can even be considered to build a new database by establishing a collaboration alliance with a local hospital to gather up breast cancer cases that were revised by radiologist.

3. To explore alternatives for feature selection. In this research effort we designed a wrapper approach using a GA to explore the feature space in the search for an optimal subset that provides the highest discriminant power to a specific classifier algorithm; other filter approaches, such as Principal Component Analysis (PCA), typically perform faster than the evolutionary method we designed and could be used for this task in future research projects.

4. To explore more ways to exploit the information of the similar cases that are retrieved from the database. The proposed model trains the classifiers with the set of retrieved similar cases, every time a new instance has to be classified. However, the classifiers could be initially trained with the complete existing database and then re-trained or modified in an appropriate way considering the retrieved cases, without the need to completely train the classifiers again.

5. To conduct experiments considering other imbalance learning approaches, such as cost-sensitive learning and ensembles for class-imbalance.

6. To explore the possibility of implementing an hybrid approach between the proposed CBR methodology and algorithms for imbalance class learning. By modifying pre-processing techniques and including cost-sensitive learning techniques within the retrieval mechanism of the model we aim to generate training sets containing not only similar cases, but also a balanced number of instances from both positive and negative class.

7. To implement *relevance feedback* on the model to enable the radiologist to evaluate the response of the CBR system and, eventually, validate the answer. This feedback would have the objective of enhancing the case-retrieval and classification mechanisms.

11.2 Summary

In this chapter we presented the conclusions that can be drawn from this doctoral dissertation. A summary of all the results obtained in the three stages of experiments that were presented previously was also included; with them we concluded that the proposed CBR model is much more accurate that the traditional CAD scheme and the classifier ensembles combined with SMOTE+ENN. Finally, we listed the potential future work that can be built upon the classification methodology that has been developed and tested in this research effort.

Bibliography

[1] Depeursinge A, Fischer B, Müller H, and Deserno TM. Prototypes for content-based image retrieval in clinical practice. *The Open Medical Informatics Journal*, 5:58–72, 2011.

[2] Agnar Aamodt and Enric Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications*, 7(1):39–59, 1994.

[3] Mobyen Uddin Ahmed, Shahina Begum, Peter Funk, Ning Xiong, and Bo von Scheele. A multi-module case-based biofeedback system for stress treatment. *Artificial Intelligence in Medicine. Special issue on Advances in Case-Based Reasoning in the Health Sciences*, 51:107–115, 2011.

[4] Mobyen Uddin Ahmed, Shahina Begum, Erik Olsson, Ning Xiong, and Peter Funk. *Successful Case-based Reasoning Applications*, chapter 2, pages 7–52. Studies in Computational Intelligence. Springer-Verlag, 2010.

[5] Edén A. Alanís-Reyes, José L. Hernández-Cruz, Jesús S. Cepeda, Camila Castro, Hugo Terashima-Marín, and Santiago E. Conant-Pablos. Analysis of machine learning techniques applied to the classification of masses and microcalcification clusters in breast cancer computer-aided detection. *Journal of Cancer Therapy*, 3(6):1020–1028, 2012.

[6] A. Andoni and P. Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *IEEE Symposium on Foundations of Computer Science*, volume 51, pages 459–468, 2006.

[7] Eva Armengol. Classification of melanomas *in situ* using knowledge discovery with explained case-based reasoning. *Artificial Intelligence in Medicine. Special issue on Advances in Case-Based Reasoning in the Health Sciences*, 51:93–105, 2011.

[8] Sunil Arya, David M. Mount, Nathan S. Netanyahu, Ruth Silverman, and Angela Y. Wu. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *Journal of the ACM*, 45(6):891–923, 1998.

[9] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[10] J. Bozek, M. Mustra, K. Delac, and M. Grgic. *A Survey of Image Processing Algorithms in Digital Mammography*, volume 231 of *Recent Advances in Multimedia Signal Processing and Communications*. Springer Berlin / Heidelberg, 2009.

[11] P. S. Bradley and O. L. Mangasarian. Massive data discrimination via linear support vector machines. *Optimization Methods and Software*, 13:1–10, 2000.

[12] L. Breiman. Bagging predictors. *Machine Learning*, 26(2):123–140, 1996.

[13] L. Breiman, J.H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth and Brooks/Cole Publishing Company, Monterey, CA, 1984.

[14] Renato Campanini and Nico Lanconelli. *Genetic Algorithms in CAD Mammography*, chapter 4, pages 129–158. Recent Advances in Breast Imaging, Mammography, and Computer-Aided Diagnosis of Breast Cancer. SPIE Press, 2006.

[15] E. Cantú-Paz. Feature subset selections, class separability and genetic algorithms. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO)*, pages 957–970, 2004.

[16] E. Cantú-Paz, S. Newsam, and C. Kamath. Feature selection in scientific applications. In *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 788–793, 2004.

[17] K. R. Castleman. *Digital Image Processing*. Prentice-Hall, Englewood Cliffs, NJ, 1996.

[18] N. V. Chawla, K. W. Bowywe, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligent Research*, 16:321–357, 2002.

[19] N.V. Chawla, N. Japkowicz, and A. Kolcz. Workshop on learning from imbalanced data sets ii. In *Proceedings of the International Conference on Machine Learning*, 2003.

[20] N.V. Chawla, N. Japkowicz, and A. Kolcz. Editorial: Special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, 6(1):1–6, 2004.

[21] H. D. Cheng, X. J. Shi, R. Min, L. M. Hu, X. P. Cai, and H. N. Du. Approaches for automated detection and classification of masses in mammograms. *Pattern Recognition*, 39(4):646–668, 2006.

[22] S. Ciatto, M. Del Turco, G. Risso, S Catarzi, R. Bonardi, V. Viterbo, P. Gnutti, B. Guglielmoni, L. Ponelli, A. Pandiscia, F. Navarra, A. Lauria, R. Palmiero, and P. Indovina. Comparison of standard reading and computer aided detection (cad) on a national proficiency test of screening mammography. *European Journal of Radiology*, 45(2):135–138, 2003.

[23] W. Cohen. Fast effective rule induction. In A. Prieditis and S. Russell, editors, *Proceedings of the 12th international conference on machine learning, ICML*, pages 115–123, Tahoe City, 1995. Morgan Kaufmann.

[24] S. E. Conant-Pablos, Rolando R. Hernández-Cisneros, and H. Terashima-Marín. *Feature Selection for the Classification of Digital Mammograms using Genetic Algorithms, Sequential Search and Class Separability.* Genetic and Evolutionary Computation: Medical Applications. S. Smith and S. Cagnoni. Wiley, 2010.

[25] M. Datar, N. Immorlica, P. Indyk, and V. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Symposium on Computational Geometry*, pages 253–262. ACM Press, 2004.

[26] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (CSUR)*, 40(2):1–60, 2008.

[27] T. Deserno. *Biomedical Image Processing*. Springer Verlag, 2011.

[28] A. R. Dominguez and A. F Nandi. Enhanced multi-level thresholding segmentation and rank based region selection for detection of masses in mammograms. *IEEE International Conference on Acoustics, Speech and Signal Processing 2007*, pages 449–452, April 2007.

[29] S. B. Edge, D. R. Byrd, C. C. Compton, A. G. Fritz, F. L. Greene, and A. Totti. *The AJCC Cancer Staging Manual*. 5th Printing. Springer, 7th edition, 2010.

[30] J. Ge, B. Sahiner, L. M. Hadjiiski, H.P. Chan, J. Wei, M. A. Helvie, and C. Zhou. *Computer-aided detection of clusters of microcalcifications on full field digital mammograms*. 2006.

[31] A. Gionis, P. Indyk, and R. Motwani. Similarity search in high dimensions via hashing. In *International Conference on Very Large Data Bases*, pages 518–529, 1999.

[32] Cai H, Peng Y, Ou C, Chen M, and Li L. Diagnosis of breast masses from dynamic contrast-enhanced and diffusion-weighted mr: A machine learning approach. *PLoS One*, 9(1), January 2014.

[33] L. Hadjiiski, B. Sahiner, H.P. Chan, N. Petrick, M.A. Helvie, and M. Gurcan. Analysis of temporal changes of mammographic features: Computer-aided classification of malignant and benign breast masses. *Medical Physics*, 28(11):2309–2317, 2001.

[34] Trevor Hastie, Rober Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, second edition, 2008.

[35] Trevor Hastie, Robert Tibshirani, and Andreas Buja. Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association*, 89(428):1255–1270, 1994.

[36] S. Haykin. *Neural Networks: A comprehensive Foundation*. Macmillan College Publishing Co., New York, second edition, 1999.

[37] Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,*, 21(9), 2009.

[38] R. R. Hernández-Cisneros and H. Terashima-Marin. Classification of individual and clustered microcalcifications in digital mammograms using evolutionary neural networks. In *MICAI 2006*, pages 1200–1210, Apizaco, Tlaxcala, Mexico, 2006. Springer Verlag.

[39] R. R. Hernández-Cisneros and H. Terashima-Marín. Evolutionary neural networks applied to the classification of microcalcification clusters in digital mammograms. In *Proceedings of the 2006 IEEE Congress on Evolutionary Computation*, pages 2459–2466, Vancouver, BC, Canada, 2006.

[40] R. R. Hernández-Cisneros and H. Terashima-Marín. Feature selection for the classification of microcalcification clusters in digital mammograms using genetic algorithms. In *GECCO Workshop on Medical Applications of Genetic and Evolutionary Computation MedGEC 2006*, Seattle, USA, 2006.

[41] R. R. Hernández-Cisneros, H. Terashima-Marín, and S. E. Conant-Pablos. Comparison of class separability, forward sequential search and genetic algorithms for feature selection in the classification of individual and clustered microcalcifications in digital mammograms. In *International Conference in Image Analysis and Recognition ICIAR 2007*, pages 911–922, Montreal, Canada, 2007. Springer Verlag.

[42] R. R. Hernández-Cisneros, H. Terashima-Marín, and S. E. Conant-Pablos. *Detection and Classification of Microcalcification Clusters in Mammograms using Evolutionary Neural Networks*. Artificial Intelligence in Healthcare. Springer Verlag, 2008.

[43] G. E. Hinton. Connectionist learning procedures. *Artificial Intelligence*, 40:185–234, September 1989.

[44] T. K. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998.

[45] Myles Hollander and Douglas A. Wolfe. *Nonparametric Statistical Methods*. John Wiley & Sons, Inc, Hoboken, NJ, 1999.

[46] Jens Hühn and Eyke Hüllermeier. Furia: an algorithm for unordered fuzzy rule induction. *Data Mining and Knwoledge Discovery*, 19:293–319, 2009.

[47] Jens Hühn and Eyke Hüllermeier. An analysis of the furia algorithm for fuzzy rule induction. *Advances in Machine Learning*, I:32–344, 2010.

[48] Alan Julian Izenman. *Modern Multivariate Statistical Techniques. Regression, Classification and Manifold Learning*. Springer, New York, USA, 2008.

[49] A. K. Jain, R. P. W. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000.

[50] N. Japkowicz. Learning from imbalanced data sets. Technical Report WS-00-05, American Association for Artificial Intelligence, 2000.

[51] Hao Jing and Yongyi Yang. Image retrieval for computer-aided diagnosis of breast cancer. In *Image Analysis Interpretation (SSIAI), 2010 IEEE Southwest Symposium on*, pages 9 –12, may 2010.

[52] Kenneth Jong. Learning with genetic algorithms: An overview. *Machine Learning*, 3(2-3):121–138, 1988.

[53] Choi JY, Kim DH, Choi SH, and Ro YM. Multiresolution local binary pattern texture analysis for false positive reduction in computerized detection of breast masses on mammograms. *SPIE Medical Imaging*, 57(21), October 2012.

[54] Felicia Marie Knaul, Gustavo Nigenda, Rafael Lozano, Hector Arreola-Ornelas, Ana Langer, and Julio Frenk. Breast cancer in mexico: a pressing priority. In Elsevier, editor, *Reproductive Health Matters*, volume 16, pages 113–123, 2008.

[55] J. L. Kolodner. Maintaining organization in a dynamic long-term memory. *Cognitive Science*, 7(4):243–280, 1983.

[56] J. L. Kolodner. Reconstructive memory: A computer model. *Cognitive Science*, 7(4):281–328, 1983.

[57] J.L. Kolodner. *Case-based reasoning*. Morgan Kauffman, San Mateo, 1993.

[58] P. Koton. *Using experience in learning and problem solving*. PhD thesis, Massachusetts Institute of Technology, Laboratory of Compute Science, 1989.

[59] V. Kurkova. *Kolmogorov's theorem*. MIT Press, Cambridge, Massachusetts, 1995.

[60] H. Li, Y. Wang, K.J.R. Liu, S.C.B. Lo, and M.T. Freedman. Computerized radiographic mass detection. part i: Lesion site selection by morphological enhancement and contextual segmentation. *IEEE Transactions on Medical Imaging*, 20(4):289–301, 2001.

[61] H.D. Li, M. Kallergi, L.P. Clarke, V.K. Jain, and R.A. Clark. Markov random field for tumor detection in digital mammography. *IEEE Transactions on Medical Imaging*, 14(3):565–576, 1995.

[62] American College of Radiology. *Breast Imaging Reporting and Data System (BI-RADS)*. American College of Radiology, Reston, VA, 1998.

[63] Imaginis Corporation. Common forms of breast cancer. `http://www.imaginis.com/breast-health/what-is-breast-cancer-1`, April 2014.

[64] Imaginis Corporation. Digital mammography. `http://www.imaginis.com/breast-health/digital-mammography-2`, April 2014.

[65] Imaginis Corporation. How is mammography performed? `http://www.imaginis.com/breast-health/general-information-on-mammography-1`, April 2014.

[66] Imaginis Corporation. How mammography is performed: Imaging and positioning. `http://www.imaginis.com/breast-health/how-mammography-is-performed-imaging-and-positioning-2`, April 2014.

[67] National Cancer Institute. Breast anatomy. `http://training.seer.cancer.gov/breast/anatomy`, April 2014.

[68] National Cancer Institute. Breast cancer screening. `http://www.cancer.gov/cancertopics/pdq/screening/breast/healthprofessional/page4`, April 2014.

[69] Salem Radiology Consulting. How is digital mammography performed? `http://www.salemradiology.com/mammography.htm`, April 2014.

[70] Salem Radiology Consulting. Views taken during screening and diagnostic mammography. `http://www.salemradiology.com/mammography.htm`, April 2014.

[71] World Health Organization. Media centre. fact sheet no. 297: Cancer. `http://www.who.int/mediacentre/factsheets/fs297/en/index.html`, April 2014.

[72] Sushmita Mitra and Tinku Acharya. *Data Mining. Multimedia, Soft Computing, and Bioinformatics*. Wiley, 2003.

[73] N.R. Mudigonda, R.M. Rangayyan, and J.E.L. Desautels. Gradient and texture analysis for the classification of mammographic masses. *IEEE Transactions on Medical Imaging*, 19(10):1032–1043, 2000.

[74] Bego na Acha, Carmen Serrano, Rangaraj M. Rangayyan, and J.E. Leo Desautels. *Detection of Microcalcifications in Mammograms*, chapter 9, pages 291–314. Recent Advances in Breast Imaging, Mammography, and Computer-Aided Diagnosis of Breast Cancer. SPIE Press, 2006.

[75] S. Oporto-Díaz, R. R. Hernández-Cisneros, and H. Terashima-Marín. Detection of microcalcification clusters in mammograms using a difference of optimized gaussian filters. In M. Kamel and A. Campilho, editors, *Proceedings of the Second International Conference on Image Analysis and Recognition, ICIAR 2005*, volume 3656, pages 998–1005, Toronto, ON, Canada, 2005. Springer, Heidelberg.

[76] Casti P, Mencattini A, Salmeri M, Ancona A, Mangieri F, and Rangayyan RM. Measures of radial correlation and trend for classification of breast masses in mammograms. In *IEEE Engineering in Medicine and Biology Society*, pages 6490–6493. IEEE, 2013.

[77] A. Papadopoulos, D.I. Fotiadis, and A. Likas. An automatic microcalcification detection system based on a hybrid neural network classifier. *Arficial Intelligence in Medicine*, 25(2):149–167, 2002.

[78] Yoon-Joo Park, Se-Hak Chun, and Byung-Chun Kim. Cost-sensitive case-based reasoning using a genetic algorithm: Application to medical diagnosis. *Artificial Intelligence in Medicine. Special issue on Advances in Case-Based Reasoning in the Health Sciences*, 51:133–145, 2011.

[79] E. Pisano, C. Gatsonis, E. Hendrick, M. Yaffe, J. Baum, S. Acharyya, E. Conant, L. Fajardo, L. Bassett, C. D'Orsi, R. Jong, and M. Rebner. Diagnostic performance of digital versus film mammography for breast cancer screening. the results of the american college of radiology imaging network (acrin) digital mammographic imaging screening trial (dmist). *New England Journal of Medicine*, October 2005.

[80] J. R. Quinlan. *Induction on decision trees*, volume 1 of *Machine Learning*. 1986.

[81] R. M. Rangayyan. *Biomedical Image Analysis*. CRC Press LLC, Boca Raton, 2005.

[82] Yvan Saeys, Iñaki Inza, and Pedro Larra naga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

[83] Farhang Sahba, Hamid R. Tizhoosh, and Magdy M. A. Salama. *Reinforced Medical Image Segmentation*, chapter XI, pages 327–346. Computational Intelligence in Medical Imaging. Techniques and Applications. CRC Press, 2009.

[84] D. Sheskin. *Handbook of parametric and nonparametric statistical procedures*. Boca Raton: Chapman & Hall/CRC, 2004.

[85] Edward A. Sickles. Mammographic features of early breast cancer. In *American Roentgen Ray Society*, 143, pages 461–464, 1984.

[86] R. L. Simpson. A computer model of case-based reasoning in problem solving: An investigation in the domain of dispute mediation. Technical report, Georgia Institute of Technology, School of Information and Computer Science, Atlanta, US, 1985.

[87] Sameer Singh and Keir Bovis. *A Weighted Gaussian Mixture Model with Markov Random Fields and Adaptive Expert Combination Strategy for Segmenting Masses in Mammograms*, chapter VIII, pages 263–289. SPIE Press, 2006.

[88] J. Suckling, J. Parker, and D. Dance. The mammographic image analysis society digital mammogram database. *Exerpta Medica. International Congress Series 1069*, pages 375–378, 1994.

[89] Yanmin Sun, Andrew K. C. Wong, and Mohamed S. Kamel. Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(4):687–719, 2009.

[90] G.M. te Brake and N. Karssemeijer. Segmentation of suspicious densities in digital mammograms. *Medical Physics*, 28(2):259–266, 2001.

[91] Bhavani Thuraisingham. *Managing and Mining Multimedia Databases*. CRC Press, 2001.

[92] S. Timp and N. Karssemeijer. Interval change analysis to improve computer aided detection in mammography. *Medical Image Analysis*, 10(1):82–95, 2006.

[93] J. T. Tou and R. C. Gonzalez. *Pattern Recognition Principles*. Addison-Wesley, London, 1974.

[94] Carol Tukington and Karen Krag. *Encyclopedia of Breast Cancer*. Facts on File Library of Health and Living, 2005.

[95] Vladimir Vapnik. *Statistical Learning Theory*. John Wiley & Sons, New York, 1998.

[96] C. Varela, P. G. Tahoces, A. J. Mendez, M. Souto, and J. J. Vidal. Computerized detection of breast masses in digitized mammograms. *Computers in Biology and Medicine*, 37:214–226, 2007.

[97] W.J.H. Veldkamp, N. Karssemeijer, J.D.M. Otten, and J.H.C.L. Hendriks. Automated classification of clustered microcalcifications into malignant and benign types. *Medical Physics*, 27(11):2600–2608, 2000.

[98] B. Verma and J. Zakos. A computer-aided diagnosis system for digital mammograms based on fuzzy-neural and feature extraction techniques. *IEEE Transactions on Information Technology in Biomedicine*, 5(1):46–54, 2001.

[99] L. Wei, Y. Yang, R.M. Nishikawa, and Y. Jiang. A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications. *IEEE Transactions on Medical Imaging*, 24(2):371–380, 2005.

[100] Liyang Wei, Yongyi Yang, and Robert M. Nishikawa. Microcalcification classification assisted by content-based image retrieval for breast cancer diagnosis. *Pattern Recognition*, 42:1126–1132, 2009.

[101] D. L. Wilson. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Communications,*, 2(3):408–421, 1972.

[102] Z. Q. Wu, Jianmin Jiang, and Y. H. Peng. *Computational Intelligence on Medical Imaging with Artificial Neural Networks*, chapter I, pages 1–26. Computational Intelligence in Medical Imaging. Techniques and Applications. CRC Press, 2009.

[103] P. Zhang, B. Verma, and K. Kumar. Neural vs. statistical classifier in conjunction with genetic algorithm based feature selection. *Pattern Recognition Letters*, 26(7):909–919, 2005.

[104] Bin Zheng. Computer-aided diagnosis in mammography using content-based image retrieval approaches: Current status and future perspectives. *Algorithms*, 2:828–849, 2009.

Vita

This doctoral dissertation was typed in using LaTeX 2_ε[1] by Edén Alejandro Alanís Reyes.

[1]The style file `phdThesisFormat.sty` used to set up this dissertation was prepared by the Center of Intelligent Systems of the Instituto Tecnológico y de Estudios Superiores de Monterrey, Monterrey Campus