



UNIVERSITÀ DEGLI STUDI DI  
CASSINO E DEL LAZIO MERIDIONALE

Corso di Dottorato in  
Metodi modelli e tecnologie per l'Ingegneria

curriculum Ingegneria dell'informazione

Ciclo XXXII

Deep Learning for computer-aided detection and diagnosis of  
clustered microcalcifications on digital mammograms

SSD:ING-INF/05

Coordinatore del Corso  
Chiar.ma Prof.ssa Wilma Polini

Dottoranda  
Benedetta Savelli

Supervisore  
Chiar.mo Prof. Claudio Marrocco

*Alla mia famiglia e a tutte le persone che mi hanno  
accompagnato in questo meraviglioso percorso di  
vita...*

# Abstract

Breast cancer is one of the most common cause of cancer death in women worldwide. In most western countries, screening programs are organized in order to detect breast cancers at an early stage. To improve breast cancer detection, many radiologists use computer-aided detection and diagnosis (CAD) systems which are able to detect and characterize mammographic signs of malignancy such as clustered microcalcifications and masses through computerized image analysis. Even though effective in terms of sensitivity, these systems produce a too high number of false alarms, which potentially limits the benefit they can provide. This thesis addresses the problem of accurately detecting and classifying clustered microcalcifications in full field digital mammograms. The goal is to reduce the gap between CAD systems and radiologists in terms of false alarms while maintaining the high sensitivity typical of the commercial CAD systems. To this end, three main contributions are proposed by exploiting innovations and advantages of novel machine learning algorithms, based on deep learning convolutional neural networks (CNNs) : (i) a new proposal of combination of a deep cascade of boosting classifiers and a CNN to deal with the high-imbalance problem of classifying individual pixels in a mammogram as belonging to a microcalcification or not; (ii) a novel method for detecting individual calcifications that provides for the use of multiple-depth CNNs, to exploit both the local features and the surrounding context of MCs; and (iii) a novel end-to-end system able to combine both detection and classification of malignant cluster, by additionally segmenting individual calcifications. Along with these contributions, experimental comparisons with other existing methods in the literature are provided and show significant reduction in the number of false alarms. Moreover a novel end-to-end model that combines detection and classification steps is presented, by showing a significant improvement with respect to single-task systems. When applied to a clinical setting, this would help the radiologists to reduce the number of unnecessarily recalled women with microcalcification clusters, thus improving the effectiveness of screening and diagnosis processes.

## ABSTRACT

---



# Contents

<b>Abstract</b>	v
<b>Summary</b>	v
<b>List of figures</b>	xi
<b>List of tables</b>	xv
<b>1 Introduction</b>	<b>1</b>
1.1 Anatomy of the breast	2
1.2 Malignant breast diseases	4
1.3 Breast imaging	6
1.3.1 Mammography	6
1.3.2 Screen-film mammography	7
1.3.3 Full-field digital mammography	8
1.4 Signs of breast cancer in mammography	9
1.4.1 Microcalcifications	9
1.4.2 Soft tissue lesions	14
1.5 Computer aided detection and diagnosis	14
1.6 Evaluation of CAD	16
1.7 Outline of the thesis	17
<b>2 Deep Learning</b>	<b>19</b>
2.1 Deep Feedforward Neural Networks	20
2.1.1 Gradient-based Learning: Stochastic Gradient Descent	21
2.1.2 Learning rate	22
2.1.3 Momentum	23
2.1.4 Dropout	23
2.2 Convolutional Neural Networks	24

## CONTENTS

---

2.2.1	Convolutional layers	24
2.2.2	Max pooling layers	25
2.2.3	Fully connected layers	26
2.3	Deep CNN architectures	27
2.3.1	Classification architectures	27
2.3.2	Segmentation architectures	29
<b>3</b>	<b>Computer aided detection of individual microcalcifications</b>	<b>33</b>
3.1	The Cascade approach	34
3.1.1	Ranking based cascade and feature set	34
3.1.2	Detection phase	34
3.1.3	Learning procedure	35
3.1.4	Deep Cascade	35
3.2	Combining Deep Cascade and Convolutional Neural Networks	36
3.2.1	Materials	37
3.2.2	Experiments	37
3.3	Results	39
3.4	Discussion	40
<b>4</b>	<b>Multi-context ensemble of CNNs for improving the automated detection of individual microcalcifications</b>	<b>41</b>
4.1	Multi-context CNN ensemble	42
4.2	Experimental analysis	44
4.2.1	Dataset	44
4.2.2	Network architecture	46
4.2.3	Training parameters	47
4.3	Results	48
4.4	Discussion and Conclusions	52
<b>5</b>	<b>Computer aided detection and diagnosis of clustered microcalcifications</b>	<b>55</b>
5.1	Multi-task learning	56
5.1.1	Hard parameter sharing	57
5.1.2	Soft parameter sharing	57
5.1.3	Mechanism underlying MTL	58
5.2	Proposed approach	59
5.2.1	Network architecture details	59

---

5.2.2 Multi-task loss . . . . .	60
5.2.3 Online Hard Example Mining . . . . .	62
5.3 Materials . . . . .	62
5.3.1 Dataset . . . . .	62
5.3.2 Groundtruth image generation . . . . .	63
5.3.3 Data samples extraction . . . . .	65
5.4 Experiments . . . . .	65
5.4.1 Performance evaluation . . . . .	65
5.4.2 Model parameters . . . . .	67
5.5 Results . . . . .	69
5.5.1 Cluster detection . . . . .	69
5.5.2 Cluster detection and classification . . . . .	69
5.6 Conclusions . . . . .	70
<b>6 Summary and Conclusions</b>	<b>75</b>
<b>Bibliography</b>	<b>79</b>

## CONTENTS

---

# List of Figures

1.1	(a) Anatomy of the breast: (1) chest wall, (2) pectoralis muscles, (3) lobules, (4) nipple, (5) areola, (6) milk ducts, (7) fat and connecting tissue, (8) skin. (b) Terminal ductal lobular unit. (c) Lobular calcifications. (d) Intraductal calcifications. Source: <a href="http://www.wikipedia.org">www.wikipedia.org</a> and <a href="http://www.radiologyassistant.nl">www.radiologyassistant.nl</a> .	3
1.2	(a) Lobular Carcinoma In Situ (LCIS): (1) normal lobular cells, (2) lobular cancer cells. (b) Different stages of cancer cells growing from the milk ducts: (I) normal cells, (II) Ductal Carcinoma In Situ (DCIS), (III) Invasive Ductal Carcinoma (IDC). Source: <a href="http://www.breastcancer.org">www.breastcancer.org</a> .	4
1.3	(a) Mammography apparatus : (1) anode, (2) filter, (3) X-rays, (4) compression plate, (5) scattering, (6) grid, (7) receptor.	8
1.4	(a)(b) Standard digital mammography exam with cranio-caudal (right) and mediolateral oblique (left) views of both breasts (source: <a href="http://TheBreastJournal.com">The Breast Journal</a> ).	9
1.5	(a) Characteristic curve of a mammographic screen-film system. (b) Characteristic response of a detector designed for digital mammography.	9
1.6	Classification of breast calcifications into benign, suspicious and malignant types basing on their distribution	12
1.7	Classification of breast calcifications into benign, suspicious and malignant types basing on their distribution	13
1.8	An example of an ROC curve	16
2.1	An example of deep feed forward neural network	20
2.2	ReLU activation function	22
2.3	Stochastic Gradient descent: the role of learning rate	23
2.4	Stochastic Gradient descent: the role of momentum	24
2.5	Comparison between a standard deep neural network and the same network with dropout application. The circles with a cross symbol inside denote deactivated units.	25
2.6	Convolutional neural network with two convolutional layers, one pooling layer and one dense layer. The activations of the last layer are the output of the network.	26
2.7	Illustration of translation invariance in convolutional neural network. The bottom leftmost input is a translated version of the upper leftmost input image by one-pixel right and one-pixel down.	27

## LIST OF FIGURES

---

2.8	VGGnet configurations. The depth of the configurations increases from the left (A) to the right (E), as more layers are added. Source: Simonyan et al. [48]	29
2.9	U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations. Source: Ronneberger et al. [56]	30
3.1	The Haar-like feature groups used by the cascade of classifiers. (a) Some examples of the first group. (b) An example of the second group. (c) Some examples of the third group.	34
3.2	Overview of the proposed MC detection scheme. On the left, Deep Cascade reduces the class imbalance ratio in the input data by about three orders of magnitude. This is achieved thanks to a sequence of five high-sensitivity classifiers that linearly combine a large number of decision stumps constrained to use single Haar-like features (top row in each classifier's box). The remaining samples are then classified by a VGGNet inspired CNN and assigned a probability score using the output from the last fully connected layer.	36
3.3	Average ROC curves obtained from 1,000 bootstrap iterations. Confidence bands indicate 95% confidence intervals along the TPR axis.	40
4.1	Overview of the proposed architecture	43
4.2	Details of the proposed architecture	45
4.3	Chart describing the findings in the INbreast database	46
4.4	Some examples of images from (a-b) INbreast	47
4.5	InBreast annotation example	48
4.6	Average ROC curves obtained from 1,000 bootstrap iterations for INbreast dataset. Confidence bands indicate 95% confidence intervals along the TPR axis.	50
4.7	FROC curves for (a) INbreast dataset	51
5.1	Hard parameter sharing for multi-task learning in deep neural networks	57
5.2	Soft parameter sharing for multi-task learning in deep neural networks	58
5.3	Overview of the proposed method	60
5.4	DIAG dataset annotation example	63
5.5	Comparison image-based FROC curves of a basic U-Net for detection and of the proposed modified U-Net for detection and classification	66
5.6	Comparison case-based FROC curves of a basic U-Net for detection and of the proposed modified U-Net for detection and classification	67
5.7	Comparison image-based ROC curves of a U-Net for classification and of the proposed modified U-Net for detection and classification	68

## LIST OF FIGURES

---

5.8	Image-based FROC curves obtained for single and joint predictions	68
5.9	Example of a True Positive detected cluster	72
5.10	Example of a False Positive detected cluster	73

## LIST OF FIGURES

---



# List of Tables

2.1	A list of commonly applied last layer activation functions for various tasks . . . . .	26
3.1	Architecture of the VGGNet-based CNN . . . . .	38
3.2	Comparative results of mean MC detection sensitivity $\bar{S}$ . . . . .	39
3.3	Average per-mammogram processing time . . . . .	39
4.1	Details of the <i>incremental block</i> . . . . .	44
4.2	Details of the <i>classification block</i> . . . . .	44
4.3	Results of mean MC detection sensitivity $\bar{S}$ for standalone CNNs . . . . .	49
4.4	Results of mean MC sensitivity $\bar{S}$ for combined CNNs . . . . .	49
4.5	Results of MC sensitivity $S$ for combined CNNs according to different combination rules . . . . .	49
4.6	Comparative results of mean MC detection sensitivity $\bar{S}$ . . . . .	51
4.7	Comparative results of the FROC score and sensitivities at specific FPP1 . . . . .	52
4.8	Results of MC per-image processing time for the trained networks . . . . .	52
5.1	Details of the <i>classification block</i> . . . . .	59
5.2	Distribution of the digital mammography (DM) exams included in this study. . . . .	62
5.3	Hyperparameter tuning and optimization . . . . .	65

## LIST OF TABLES

---

# Chapter 1

## Introduction

Worldwide, breast cancer is the most common cancer (24.2%) and the first known cause of death (13,7%) among women aged between 35 and 55 [1]. Detecting breast cancer as early as possible is vital to improve patient's chances and quality of life after treatment. With this aim, population-based screening program started in the late 70s and have been adopted as organized nation-wide programs in many developed countries since then. In the screening programs asymptomatic women within a certain age range are regularly invited (every year or every two years) to obtain a screening exam. It is important to underline that the positive effects of screening are mainly due to the principle of repetition. For this reason, the first round should be considered differently from the repeated rounds and monitored and reported separately.

The chosen technique for screening mammography, is a noninvasive and relatively cheap test which uses x-rays to obtain a two-dimensional(2D) image of the breast. Although using ionizing radiation, the risk for an average 50 year old woman to develop cancer from a mammography exam is estimated to be about 9 in 10000 [2]. Screen film mammography was initially used, until it was replaced by digital mammography (DM) in the mid-2000s, showing an improvement in breast cancer detection accuracy, especially for women with dense breasts [3].

The large number of acquired screening mammograms are interpreted by radiologists, who look for mammographic indicators of cancer like clusters of microcalcifications (MCs) and masses, and subsequently make a final decision whether the woman has to be recalled for further assessment. However, interpreting screening mammograms is a big challenge even for an expert radiologist since the low prevalence makes finding abnormalities difficult. In [4] are pointed out several subjective factors that may lead to a lack of perception or to mistakes in interpretation. Among the established methods to improve radiologist performance, it has been reported that having more than one radiologist or a CAD system improves the detection of cancer in mammograms [5, 6]. It is common practice that each woman' screening exam is reviewed by two readers, in an independent double reading fashion. If the two radiologists disagree, the final decision on the need to recall the woman can be either by consensus of the two radiologists, by arbitration by a third radiologist, or the woman can be recalled if either of the two radiologists decides that a recall is warranted. If recalled, the woman is referred to go to a hospital for further tests (diagnostic work-up). Double reading and therefore combining assessments by two or more readers improves overall

performance. However several studies have shown that unfortunately up to 25% of mammography detectable cancers are still missed at screening even after double reading [7, 8]. Consequently in the last few decades, Computer-Aided Detection and Diagnosis (CADe/CADx) systems have been proposed to assist radiologists in finding and locating abnormalities on the images and supporting their diagnosis response [9, 10]. To this end, several commercial CAD systems are nowadays available and their use is widespread among radiologists. However, even though CAD systems show a sensitivity similar to radiologists [11], there are still a few hundred false positives for every true positive in a screening setting, which is about two orders of magnitude higher than what the radiologists achieve [5]. This can potentially limit the benefit that a CAD system can provide, for example by resulting in an increase of the recall rate [5] and subsequently of the false positive case (i.e., patients that are recalled unnecessary), thus causing unnecessary anxiety and thereby discouraging women to participate to screening and generating lack of trust of the readers towards CAD [12]. Nevertheless, the recent developments in Artificial Intelligence techniques for Computer Vision tasks, in particular Deep Neural Networks, have brought their positive effects also in the medical image field, showing to be very effective for medical image analysis tasks [13, 14, 15]. As a consequence a new generation of CADe/CADx has been enabled with new solutions and perspective for digital mammograms tools [16, 17].

The objective of the studies described in this thesis is to develop a full CAD system for the detection and diagnosis of clusters of MCs, able to reduce the gap between CAD systems and radiologists in terms of false positives while maintaining the high sensitivity typical of commercial CAD systems. In this way, the effectiveness of CAD in assisting radiologists in screening could be improved to avoid unnecessary and invasive further work-ups in healthy women as it still happens nowadays. In this chapter an overview of the framework in which this research has been carried out is provided. Starting from a short description of the breast anatomy and the breast diseases, particular emphasis is given to the presence of suspicious and malignant MCs as one of the most important early indicator of breast cancer in mammography. Subsequently, a more detailed description of CAD system and evaluation metrics is given. Finally, an outline of the thesis is presented.

## 1.1 Anatomy of the breast

Anatomically the breast can be subdivided into the following structural entities [18, 19, 20]:

### **Chest wall**

The boundary of the thoracic cavity (see Fig. 1.1a-1).

### **Pectoralis muscles**

Thick, fan-shaped muscles, situated at the chest (anterior) of the human body (see Fig. 1.1a-2).

### **Lobules**

The basic functional unit in the breast is the *lobule*, also called the tdlu (see Fig. 1.1a-3). The tdlu consists of 10-100 *acini*, that drain

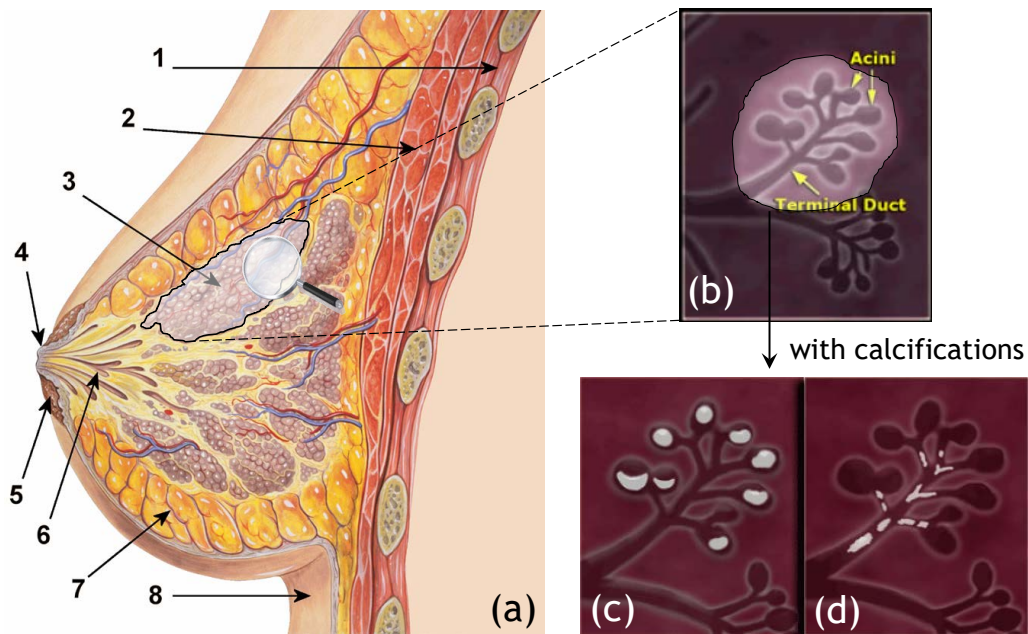


Figure 1.1: (a) Anatomy of the breast: (1) chest wall, (2) pectoralis muscles, (3) lobules, (4) nipple, (5) areola, (6) milk ducts, (7) fat and connecting tissue, (8) skin. (b) Terminal ductal lobular unit. (c) Lobular calcifications. (d) Intraductal calcifications. Source: [www.wikipedia.org](http://www.wikipedia.org) and [www.radiologyassistant.nl](http://www.radiologyassistant.nl).

into the *terminal duct* (see Fig. 1.1b). The terminal duct drains into larger ducts and finally into the main duct of the *lobe* (or segment), that drains into the *nipple*. The breast contains 15-18 lobes, each containing 20-40 lobules. The tdlu is an important structure because most invasive cancers arise from the tdlu. It is also the site of origin of dcis, lcis, fibroadenoma and fibrocystic disease, like cysts, apocrine metaplasia, adenosis and epitheliosis. Most calcifications in the breast form either within the acini (lobular calcifications, see Fig. 1.1c) or within the terminal ducts (intraductal calcifications, see Fig. 1.1d).

### Nipple

A small projection of skin containing the outlets for 15-20 lactiferous ducts arranged cylindrically around the tip (see Fig. 1.1a-4). The skin of the nipple is rich in a supply of special nerves that are sensitive to certain stimuli. The physiological purpose of nipples is to deliver milk to the infant, produced in the female mammary glands during lactation.

### Areola

Pigmented area around the nipple (see Fig. 1.1a-5).

### Milk duct

Milk ducts (or lactiferous ducts) form a tree branched system connecting the lobules of the mammary gland to the tip of the nipple (see Fig. 1.1a-6). They are the structures which carry milk toward the nipple in a lactating female.

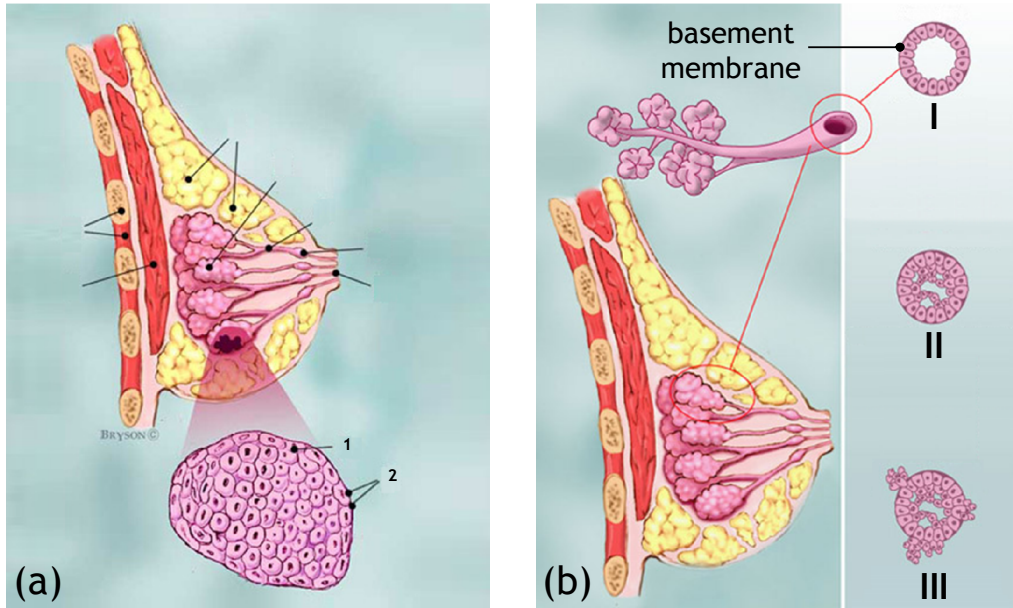


Figure 1.2: (a) Lobular Carcinoma In Situ (LCIS): (1) normal lobular cells, (2) lobular cancer cells. (b) Different stages of cancer cells growing from the milk ducts: (I) normal cells, (II) Ductal Carcinoma In Situ (DCIS), (III) Invasive Ductal Carcinoma (IDC). Source: [www.breastcancer.org](http://www.breastcancer.org).

### Fat, ligaments and connective tissue

Spaces around the lobules and ducts are filled with fat, ligaments and connective tissue (see Fig. 1.1a-7). The amount of fat in the breast largely determines their size. The actual milk-producing structures are nearly the same in all women. Female breast tissue is also sensitive to cyclic changes in hormone levels. Younger women might have denser and less fatty breast tissue than do older women who have gone through menopause.

## 1.2 Malignant breast diseases

Malignancy can grow within all types of breast tissue, but in the classical sense breast cancer originates in either the milk-ducts, in which breast milk is transported to the nipple, or the lobules, where breast milk is produced. Several types of breast cancer can arise in other parts of the breast as well, but are less common (< %8 of all breast cancers). The basement membrane (see Fig. 1.2) plays a key role in determining whether a carcinoma is “in situ” (i.e., it has not grown through the basement membrane) or “invasive” (i.e., it has grown through the basement membrane). When in situ carcinomas develop into invasive cancers, they can form metastases to lymph nodes and other organs which will decrease the survival chance. In the breast the following forms of malignancy can be considered [21]:

### Lobular Carcinoma In Situ (LCIS)

LCIS is an area (or areas) of abnormal cell growth in the lobule (see Fig. 1.2a). The abnormal cells start growing in the lobules and remain inside the lobule without spreading to surrounding tissues. People



diagnosed with LCIS tend to have more than one lobule affected. Despite the fact that its name includes the term “carcinoma”, LCIS is not a true breast cancer. Rather, LCIS is an indication that a person is at higher-than-average risk for getting breast cancer at some point in the future.

LCIS is usually diagnosed before menopause, most often between the ages of 40 and 50. Less than 10% of women diagnosed with LCIS have already gone through menopause. LCIS does not cause symptoms and usually does not show up on a mammogram. It tends to be diagnosed as a result of a biopsy performed on the breast for some other reason.

### **Invasive lobular Carcinoma(ILC)**

ILC is the second most common type of breast cancer after Invasive Ductal Carcinoma (IDC). The cancer begins in the milk-producing lobules and breaks through the wall of the lobule thus invading the tissues of the breast. Over time, ILC can spread to the lymph nodes and possibly to other areas of the body. About 10% of all invasive breast cancers are invasive lobular carcinomas. ILC tends to occur later in life than Invasive Ductal Carcinoma (IDC): the early 60s as opposed to the mid to late 50s. At first, ILC may not cause any symptoms. Sometimes, an abnormal area turns up on a screening mammogram, which leads to further testing. ILC tend to be more difficult to see on mammograms than IDC are. That is because instead of forming a lump, the cancer cells more typically spread to the surrounding connective tissue in a line formation.

### **Ductal Carcinoma In Situ (DCIS)**

DCIS is the most common type of non-invasive breast cancer and it represents the 25-30% of all reported breast cancers [20]. The cancer starts inside the milk ducts and remain in the ducts without spreading to the surrounding tissues (see Fig. 1.2b-II). DCIS is not life-threatening, but having DCIS can increase the risk of developing an invasive breast cancer later on. When a woman has had DCIS, she is at higher risk for the cancer coming back or for developing a new breast cancer than a woman who has never had breast cancer before. Most recurrences happen within the 5 to 10 years after diagnosis. The chances of a recurrence are under 30%. DCIS generally has no signs or symptoms. A small number of women may have a lump in the breast or some discharge coming out of the nipple. However, approximately 95% of all DCIS is diagnosed because of mammographically detected microcalcifications [20], making it the most easily detectable cancer in mammography among the early stages of cancer.

### **Invasive Ductal Carcinoma(IDC)**

IDC is the most common type of breast cancer. About 80% of all breast cancers are IDC. The cancer starts inside the milk ducts and breaks through the wall of the duct thus invading the tissues of the breast (see Fig. 1.2b-III). Over time, IDC can spread to the lymph nodes and possibly to other areas of the body. Although IDC can affect women at any age, it is more common as women grow older. According to the American Cancer Society, about two-thirds of women are 55 or older when they are diagnosed with an IDC. At first, IDC

may not cause any symptoms. Often, an abnormal area (*mass*) turns up on a screening mammogram, which leads to further testing.

### 1.3 Breast imaging

Imaging of the breast is currently done using either X-ray (mammography, tomosynthesis, CT), sound waves or radio waves:

- Mammography: Mammography involves exposing the breast to a small dose of ionizing radiation. The breast is placed in a C ark between an X-ray source emitting radiation and a detector.
- Tomosynthesis: Similar to mammography, tomosynthesis is based on X-ray. In this case the x-ray tube moves in an arc over the compressed breast, by capturing multiple images of each breast from different angles. The digital images are then reconstructed or “synthesized” into a set of three-dimensional images to get a better view of structures that would otherwise be hidden. The dose of radiation is slightly higher though within the limits of safe radiation outlined by the FDA.
- Breast CT: Similar to mammography and tomosynthesis, breast CT is based on X-ray, but instead images are taken from many different angles so as to create a full 3D reconstruction of the breast, where voxels have a quantitative meaning. Breast CT still has limited application in the clinic.
- Breast Ultrasound : Ultrasound devices use soundwaves to produce an image of the internal structure of the breast. Ultrasound is typically used as a complementary modality to mammography to diagnose lumps that were found suspicious on mammogram. Ultrasound can not look as deep inside the breast as mammography can, does not image the whole breast at once and can not see all indications (such as calcifications) that are visible on a mammogram. Is therefore unsuitable for stand-alone imaging.
- Magnetic Resonance Imaging: MRI uses magnetic fields and radiowaves to generate images of internal structure. Similar to breast ultrasound, calcifications in the breast are typically not visible in MRI. It is often used as complementary to a mammogram for women in high risk populations. The sensitivity of readers is substantially higher, but MRI is also substantially more expensive than mammography and in general with lower specificity.

#### 1.3.1 Mammography

Mammography is the oldest and still most common breast imaging technique that is used to detect and characterize breast cancer thanks to its high performance and low costs [22, 5, 23] and are used both for screening and diagnostic purposes. In general, screening mammography is performed on asymptomatic women to identify suspicious signs at an early, and therefore more treatable stage. Diagnostic mammography is performed on symptomatic patients, or to work up abnormalities found on screening images.



Hence, the aim is to characterize the pathology and define a diagnosis. In a standard examination, two images of each breast are taken: one from the top, called craniocaudal (CC) and one with the X-ray tube angled approximately  $45^\circ$  medially, called mediolateral oblique (MLO). This ensures that the images display as much breast tissue as possible. An overview of the mammography apparatus is given in Fig. 1.3. X-ray photons are emitted from the anode that is located in the X-ray tube on the top of the machine. Whereas most x-ray tubes use tungsten as the anode material, mammography equipment uses molybdenum anodes or in some designs, a dual material anode with an additional rhodium track. These materials are used because they produce a characteristic radiation spectrum that is close to optimum for breast imaging. After x-ray photons are emitted, they pass through a molybdenum (or rhodium) filter to reduce unnecessary exposure to the patient and also to improve contrast sensitivity. Part of the radiation then goes through the breast, which is compressed primarily to spread the breast tissue laterally in order to minimize the likelihood of occult cancers, and secondarily to reduce the thickness of the breast thus obtaining a clear x-ray image. As a result of interaction between breast tissue and radiation, X-ray photons may undergo a change in direction before hitting the receptor. By positioning a grid in front of the receptor, influence of scatter is reduced. Both film/screen and digital receptors are used for mammography, thus obtaining Screen-Film Mammography (SFM) and Full-Field Digital Mammography (FFDM) (see Fig. 1.3c), respectively, whose characteristics are detailed in the following.

### 1.3.2 Screen-film mammography

In screen-film based mammography, the radiation is absorbed by a scintillator (the screen) that transfers the incident X-ray photons into light photons that blacken the film, which is located just in front of the screen. The film serves as the media for recording within the receptor, transporting and storing images, and is the image display device. The significant characteristic is that the contrast of the image is “fixed” and cannot be changed after the film is exposed and chemically processed. In addition, the relation between optical film density and exposure values is non-linear (see Fig. 1.5a) and depends on the type of film used. For these reasons, it has been shown that although SFM has high sensitivity and specificity, it has some important limitations as well [24, Chapter 1]. The film used to capture, store and display the mammographic image is one of the major technical restrictions of SFM. The visibility of breast cancer depends on different attenuation of the X-ray beam by the suspect regions compared with the surrounding tissue. Suspect regions lying in dense areas of the breast may not be noticeable because film contrast decreases in the densest breast areas. This is due to limited dynamic range of conventional films. Furthermore, the image data obtained using a SFM-based system cannot be manipulated once the image is processed in a film processor. Specifically, over- and under-exposed images have to be recorded again. Contrast levels in the image cannot be altered to improve the relative visibility of structures in the image without recording additional images of the patient. Most of the several technical limitations associated with SFM are overcome by FFDM, which is described in the following section.

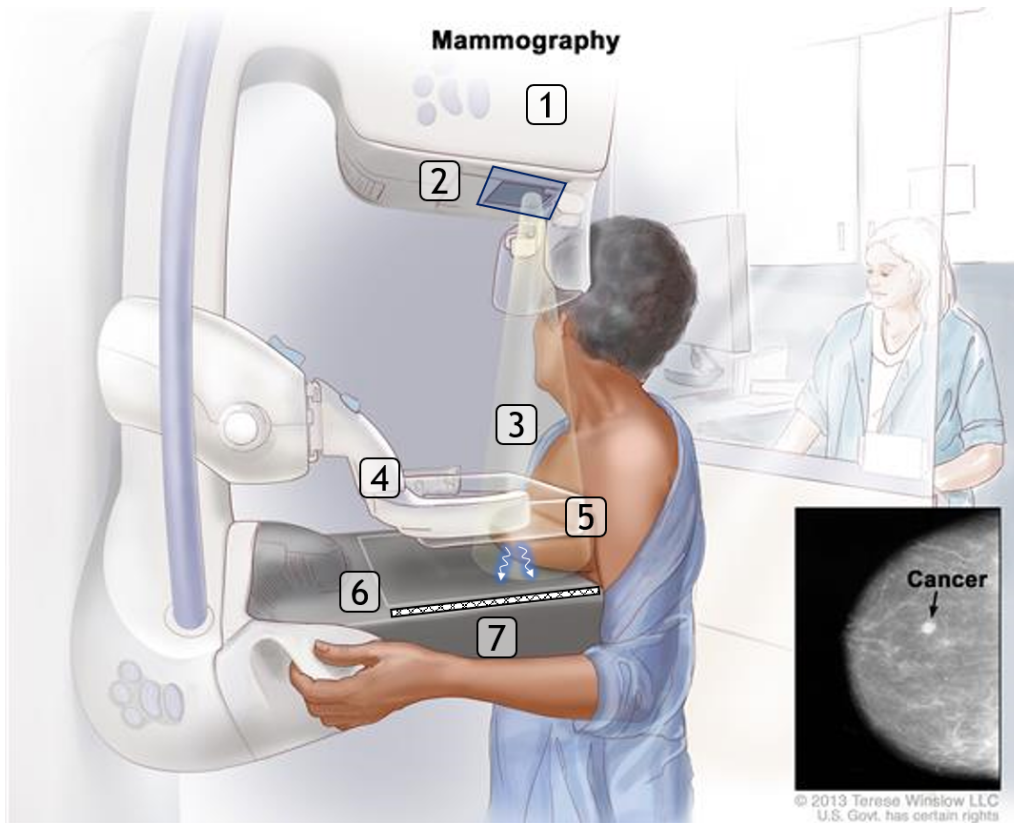


Figure 1.3: (a) Mammography apparatus : (1) anode, (2) filter, (3) X-rays, (4) compression plate, (5) scattering, (6) grid, (7) receptor.

### 1.3.3 Full-field digital mammography

In the last few years, FFDM-based systems have been developed and are increasingly used in clinical practice. In FFDM, the radiation is captured by a digital detector that inherently produces a signal that is linearly proportional to the intensity of X-rays transmitted by the breast (see Fig. 1.5). There are several key features of FFDM that distinguish it from SFM and contribute to its potential advantages [24, Chapter 1]. First of all FFDM decouples the processes of image acquisition from the subsequent stages of archiving, retrieval, image display and digital image processing. Unlike the situation in SFM where these processes are inextricably linked, this facilitates optimization of each of the separate functions and great flexibility in the adjustment of image display characteristics. Secondly FFDM it is often possible to design detectors that allow efficient use of the incident X-rays without excessive loss of spatial resolution and signal to-noise ratio. This permits a substantial reduction in the radiation dose to the breast when compared with SFM without sacrifice of image quality. Moreover, because of the differences in technology between SFM and FFDM, the optimum exposure conditions may shift toward the use of higher energy spectra than would be used with film, particularly for dense or thick breasts. In an SFM-based system, the relation between optical film density and exposure values is highly non linear and it tends to flatten for exposures above and below a fairly restricted range (see Fig. 1.5a). This limited range has important implications on image quality [24, Chapter 1]. On the contrary, in FFDM

## 1.4 Signs of breast cancer in mammography

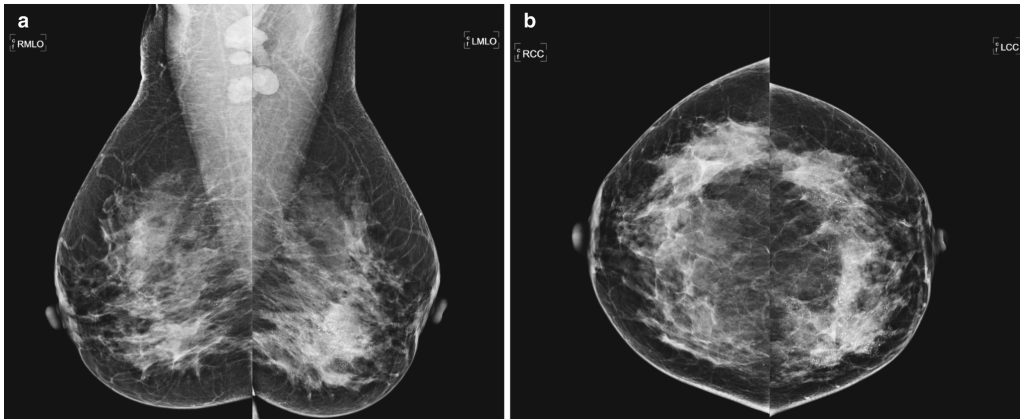


Figure 1.4: (a)(b) Standard digital mammography exam with craniocaudal (right) and mediolateral oblique (left) views of both breasts (source: [The Breast Journal](#)).

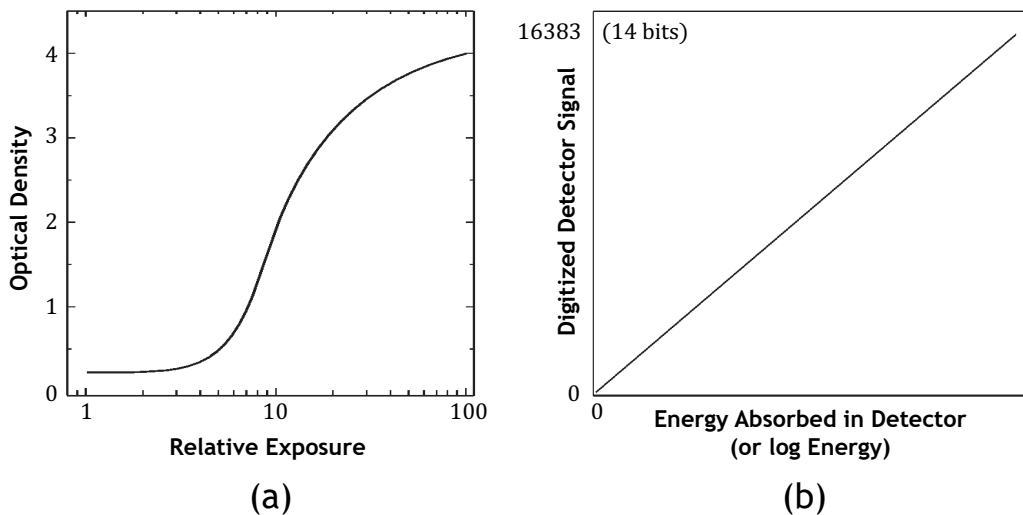


Figure 1.5: (a) Characteristic curve of a mammographic screen-film system. (b) Characteristic response of a detector designed for digital mammography.

the detector inherently produces a signal that is linearly proportional to the intensity of X-rays transmitted by the breast (see Fig. 1.5b). It has a very large dynamic range, so that it is possible to produce a faithful representation of X-ray transmission for all parts of the breast.

## 1.4 Signs of breast cancer in mammography

The main signs of malignancy on mammography can roughly be divided into two groups: microcalcifications and soft tissue lesions<sup>[25]</sup>.

### 1.4.1 Microcalcifications

Microcalcifications are calcium deposits that appear as small white specks on the mammogram. Typical size of a MCs is between 0.1 mm and 1 mm

[26] and they may appear scattered over the whole breast or distributed in one or more clusters. Microcalcification clusters may appear in both in situ and invasive breast cancer but also in benign diseases. Many of the breast cancers that are at an early stage are currently detected by the presence of microcalcifications. Approximately 95% of all DCIS is diagnosed because of mammographically detected microcalcifications [20].

Most calcifications in the breast form either within the TLDU (intraductal calcifications) or within the acini (lobular calcifications) [20]. More details on these two types of calcifications are given in the following.

### Lobular calcifications

These calcifications fill the acini, which are often dilated (see Fig. 1.1c). This results in uniform, homogeneous and sharply outlined calcifications, that are often punctate or round. When the acini become very large, as in cystic hyperplasia, “milk of calcium” may fill these cavities. However when there is more fibrosis, as in sclerosing adenosis, the calcifications are usually smaller and less uniform. In these cases it can be difficult to differentiate them from intraductal calcifications. Lobular calcifications usually have a diffuse or scattered distribution, since most of the breast is involved in the process that forms the calcifications. Lobular calcifications are almost always benign.

### Intraductal calcifications

These calcifications are calcified cellular debris or secretions within the intraductal lumen (see Fig. 1.1d). The uneven calcification of the cellular debris explains the fragmentation and irregular contours of the calcifications. These calcifications are extremely variable in size, density and form (i.e., pleomorphic from the Greek pleion “more” and morphe “form”). Sometimes they form a complete cast of the ductal lumen. This explains why they often have a fine linear or branching form and distribution. Intraductal calcifications are suspicious of malignancy.

The diagnostic approach to breast calcifications is to analyze the morphology, distribution and sometimes change over time.

### Morphology

The form or morphology of calcifications is the most important factor in deciding whether calcifications are typically benign or not. If not, they are either suspicious (intermediate concern) or of a high probability of malignancy. Usually biopsy in these cases is needed to determine the etiology of these calcifications. Using morphology as classification criterion, we can distinguish microcalcifications as follows:

- Skin calcifications: these are usually lucent-centered deposits. Skin calcifications may simulate parenchymal breast calcifications and may look like malignant-type calcifications, but when looking at MLO and CC views these calcifications look exactly the same.
- Vascular calcifications: These are linear or form parallel tracks, that are usually clearly associated with blood vessels. They may simulate intraductal calcifications.

## 1.4 Signs of breast cancer in mammography

---

- Popcorn-like calcifications: These calcifications are produced by involuting fibroadenomas. They usually do not cause a diagnostic problem.
- Rod-like calcifications: These benign calcifications form continuous rods that may occasionally be branching. They are different from malignant-type fine branching calcifications, because they are usually  $> 1mm$  in diameter. They may have lucent centers if the calcium is in the wall of the duct. These calcifications follow a ductal distribution, radiating toward the nipple and are usually bilateral.
- Round and punctuate calcifications: Round calcifications are  $0.5 - 1mm$  in size and frequently form in the acini of the terminal duct lobular unit. When smaller than  $0.5mm$ , the term punctuate is used.
- Milk of Calcium: These are benign sedimented calcifications. On CC views they appear as fuzzy, round or amorphous whereas on MLO view they may appear as semilunar crescent shaped.
- Coarse irregular lava-shaped: These calcifications are larger than  $0.5mm$  and often have a lucent center. They are seen in irradiated breast or following trauma. They develop 3 – 5 years after treatment in about 30% of women. These calcifications are also described as fat necrosis.
- Amorphous calcifications: Amorphous or indistinct calcifications are defined as without a clearly defined shape or form. These calcifications are usually so small or hazy in appearance, that a more specific morphologic classification cannot be determined.
- Coarse heterogeneous: Coarse heterogeneous microcalcifications, formerly called coarse granular, are irregular, conspicuous calcifications that are generally larger than  $0.5mm$ . They are considered to be of intermediate concern, along with amorphous microcalcifications.
- Fine pleomorphic microcalcifications: These calcifications vary in size and shapes. They are more conspicuous than the amorphous calcifications. There is a 25 – 40% risk of malignancy.
- Fine linear branching: These are thin, linear or curvilinear irregular calcifications. They may be discontinuous and their appearance suggests filling of the lumen of a duct. They have a high probability of malignancy

### Distribution

Based on the distribution calcifications can be classified as see Figs.( [1.6](#) [1.7](#)):

- Diffuse or Scattered: diffuse calcifications may be scattered calcifications or multiple appearing throughout the whole breast. It is typically seen in benign entities.
- Regional: scattered in a larger volume ( $> 2cc$ ) and not in the expected ductal distribution. Such a distribution is considered a non ductal distribution, which means associated with a benign entity.

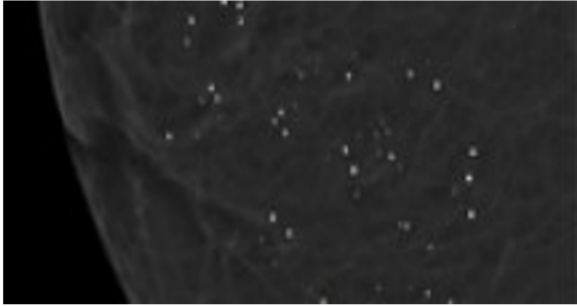
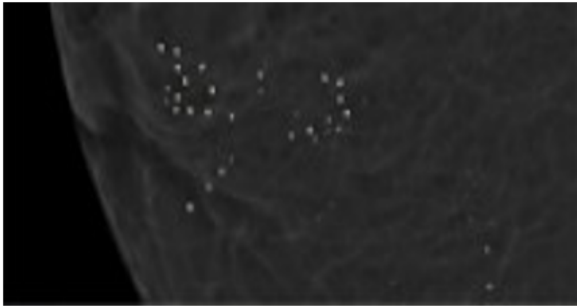

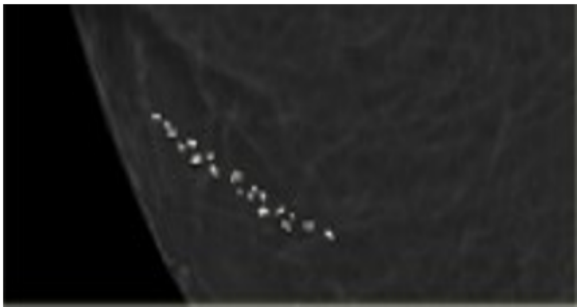
Property	Type	
distribution ----- diffuse	<b>benign</b>	
distribution ----- regional	<b>benign</b>	
distribution ----- clustered	<b>suspicious</b>	
distribution ----- linear	<b>malignant</b>	

Figure 1.6: Classification of breast calcifications into benign, suspicious and malignant types basing on their distribution



## 1.4 Signs of breast cancer in mammography

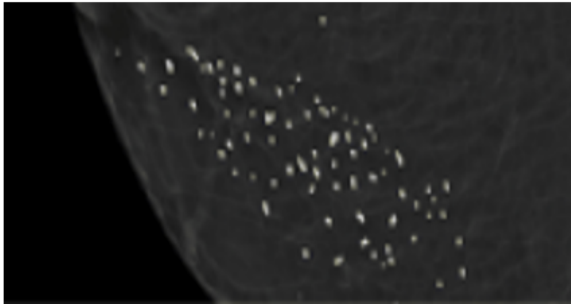
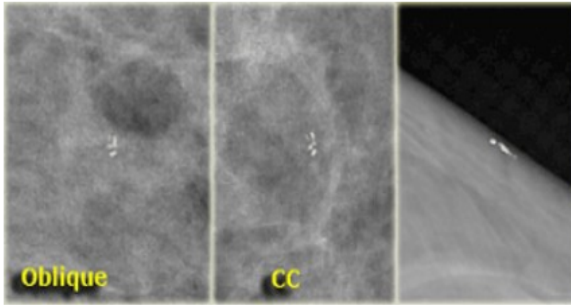
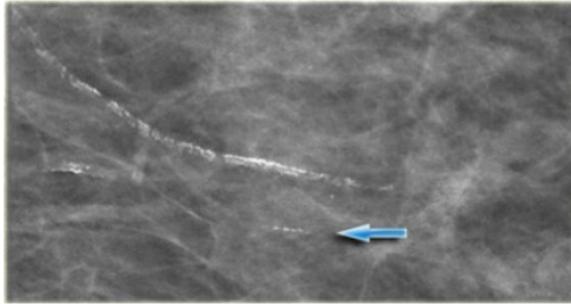
Property	Type	
distribution ----- segmental	<b>malignant</b>	
distribution ----- tattoo	<b>benign</b>	
distribution ----- vascular	<b>benign</b>	

Figure 1.7: Classification of breast calcifications into benign, suspicious and malignant types basing on their distribution

- **Clustered:** at least 5 calcifications occupying a small volume of breast tissue. Clustered microcalcifications are seen either in benign or malignant disease and are of intermediate concern. When several clusters are scattered throughout the breast tissue, this is usually considered a sign of a benign entity. On the contrary a single isolated cluster favors a malignant entity.
- **Linear:** calcifications arrayed in a line, suggesting deposits in a whole duct. This kind of distribution appears when DCIS fills the entire duct and its branches with calcifications.
- **Segmental:** calcium deposits in ducts and branches of a segment or lobe. It usually favors a ductal distribution, i.e. malignancy.

### Change over time

There are conflicting data concerning the value of absence of changeover time. It is said that the absence of interval change in microcalcifications that are probably benign on the basis of morphologic criteria is a reassuring sign and an indication for continued mammographic follow-up. On the other hand in a retrospective study that included indeterminate and suspicious clusters of microcalcifications, stability can not be relied on as a reassuring sign of benignancy.

### 1.4.2 Soft tissue lesions

When DCIS develops into an invasive cancer, the breast cancer also becomes a soft-tissue lesion, which is the term for masses, architectural distortions and asymmetrical densities within the breast. Most soft tissue lesions have the main appearance of masses and consist of cancer cells that are more densely packed together and invades the surrounding tissue, which consists mainly of fat cells and fibrous tissue. The boundary of this type of lesion can vary between circumscribed, indistinct or spiculated. The latter type, are stellated patterns of lines that are directed towards the center of the mass. These spiculations are an important sign for malignancy of the lesion. Architectural distortions are a disruption of the normal pattern in the breast without a visible mass and are less often an invasive cancer. The asymmetrical densities, a mismatch between the density pattern between the left and right breast or acquisitions at different view angles of the breast, are also less often a malignancy.

## 1.5 Computer aided detection and diagnosis

The advancement of medical imaging over the past decades resulted in a big amount of medical images, with a substantial increasing of the workload of radiologists. This is particularly true for screening programs, such as breast cancer screening, where millions of medical images are acquired each year [27, 28]. Besides the increasing workload, manual interpretation of medical images is subjective to the individual skills of radiologist and also depends on experience and their compliance with reporting guidelines.



For consistent reporting, different reporting systems are available for diseases and modalities such as the BI-RADS for breast imaging [29]. The difference in reading quality between radiologists can result in a difference in the diagnosis of a patient and can have a big impact on the number of detected cancers. For instance, a sensitivity difference varying between 18% and 40% has been observed when mammograms are read by individual breast cancer screening radiologists [30, 31]. Therefore, in many European countries, double reading has been introduced to reduce the variability in breast cancer screening performance and to increase sensitivity. However, double reading demands additional radiologists which increases their workload even more and increases costs. Moreover the low prevalence of exams with cancer (approximately 10 per thousand women screened) within the total amount of screening examinations in fact decreases radiologists' sensitivity, and therefore even double reading might not be enough. Several studies have shown that unfortunately up to 25% of mammographically detectable cancers are still missed at screening even after double reading [8]. To reduce the radiologists workload and to improve quality of reading medical images, computer-aided detection and diagnosis systems have been developed and have been extensively explored for the past decades [32, 33, 34, 10, 35, 36, 37, 38, 39]. In these systems, various automatic algorithms are used to analyse medical images and give a response to aid the radiologists. In general, there are two types of responses and, consequently, two types of CAD systems. In a CADE (computer-aided detection) system, the general aim of the system is to detect abnormalities in medical images. Therefore, the output of a CADE system are marks (or findings) of potential locations of abnormalities within the image. This type of system is mainly used to reduce the number of abnormal regions that could potentially be overlooked by the radiologist. Additionally, many of these systems supply a score with the supplied findings to show how certain the system is about a specific location to be abnormal. The second type of CAD systems are computer-aided diagnosis (CADx) systems. These systems are developed to be an aid for the radiologist in the interpretation of abnormal regions. For example, CADx systems can help in the interpretation and classification of benign and malignant disease in various diseases and imaging modalities such as breast cancer in mammography. The implementation of a CAD system into the daily workflow of radiologists can be done with different setups. For instance, a CAD system can be leveraged directly from the radiologists in reading medical images. In this setup, the aid of the system can be either aimed at the detection of abnormalities (CADE) or as an interactive decision support for the evaluation of found abnormalities (CADx). In the first scenario, CADE findings can be prompted on the image when desired to check if certain regions were not overlooked, whereas in a CADx perspective they can be shown when the radiologist wants to know if a certain region is found to be suspicious by the system. In another setup, a CAD system can be used as a completely independent reader of medical images. When used as a first reader, a possibility is the automatic preselection of mammograms based on an exam-based score denoting the likelihood that cancer is present. Furthermore the system can be a substitute of one radiologist in double reading.

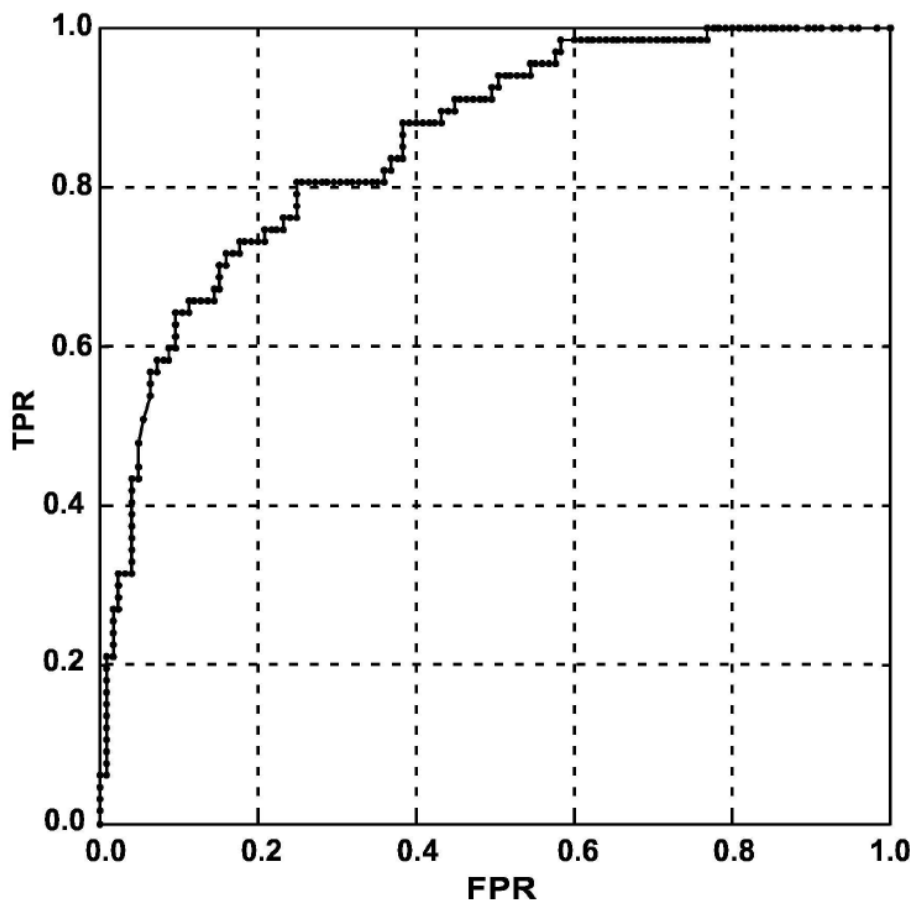


Figure 1.8: An example of an ROC curve

## 1.6 Evaluation of CAD

When a new CAD system is developed, its performance need to be validated. The validation strategy should be done as properly as possible to be able to compare the performance of the new proposed system to other CAD systems. In this section, the evaluation methods are described which are used throughout this thesis. Commonly, a CAD system is validated on a reference dataset (or ground truth) that is created based on the diagnostic findings of radiologists and the histopathological findings after diagnostic follow up. Based on these findings, annotations are drawn by capturing the malignant lesion (i.e. calcifications or soft-tissue lesions) in the image. The CAD system is applied on this dataset and the percentage of detected malignancies (the true positive rate or sensitivity) and the percentage of detected normals (the false positive rate or 1 - specificity) are calculated. Because the findings produced by the CAD system have a classification score, Receiver Operating Characteristics (ROC) analysis can be performed. With ROC analysis, all samples in the dataset are ranked according to their classification score. To obtain an ROC curve, various thresholds ( $T_h$ ) are set on the classification scores and at each threshold the number of true positives (detected malignancies) and false positives (detected non-malignancies) are

calculated to determine the operating point, i.e. the combination of the sensitivity and specificity at a given  $T_h$ . When various thresholds are set, various operating points can be calculated and an ROC curve can be plotted: an example of an ROC curve is shown in Figure 1.8. Often the Area Under the ROC Curve (AUC) is calculated to give an overall metric of the performance. The value of the AUC lies in the range of 0 and 1 where an AUC of 1 means perfect performance, i.e. all malignancies are detected without any false positive. However, to obtain a ROC curve it should be specified clearly when a finding of the CAD system is a true positive or a false positive. Moreover, when a certain range is of interest, e.g. the high specificity range between 0.8 and 1.0, the partial AUC (pAUC) can be evaluated. Furthermore, the mean sensitivity of the ROC curve in the specificity range on a logarithmic scale can be evaluated. The mean sensitivity is defined as in [40]:

$$\bar{S}(a, b) = \frac{1}{\ln(b) - \ln(a)} \int_a^b \frac{s(f)}{f} df \quad (1.1)$$

where  $a$  and  $b$  are the lower and upper bound of the false positive fraction and  $s(f)$  is the sensitivity at the false positive fraction  $f$ . Another analysis that can give a good insight in the performance of a CAD system is the Free-response ROC (FROC). Similar to ROC analysis, the number of true positives and false positives are calculated at various classification scores and the definitions are the same. However, instead of plotting the sensitivity in terms of the specificity, it is plotted in terms of the number of false positives per (normal) image (FP/I). To obtain the FROC curve,  $T_h$  is set at various values and for each value the number of false positives is determined and divided by the total number of normal images in the test set. Calculating these operating points for a FROC curve makes it possible to see how the CAD system would fit in a clinical environment because it directly show the number of false positive marks generated at a certain sensitivity. Besides directly comparing ROC (FROC) curves between different systems, a statistical comparison is also relevant for evaluation. To compare two systems bootstrapping can be used [41]. Bootstrapping is a non-parametric method to obtain confidence intervals for each curve. The bootstrapping method consists of resampling reference dataset with replacement  $n$ -times (commonly,  $n > 100$ ), and for each sample set performance metrics are evaluated for each system and compared. Statistical significance levels can be set, as for example the p-value that is defined as the fraction of performance measure values that are negative or zero, corresponding to cases in which the method performs worse or equally than the method under comparison. Hence the lower the p-value, the more statistically significant the measured performance difference. In general, it is assumed that two systems are statistical different at a  $p < 0.05$ .

## 1.7 Outline of the thesis

The final goal of the research activity described in this thesis was to develop a full CAD system for the detection and diagnosis of suspicious and malignant clusters of MCs in FFDM, with the aim of reducing the gap between CAD and radiologists in terms of FPs, while maintaining the high

sensitivity of state-of-the-art CAD commercial system. This is achieved by exploiting many of the most recent advances in deep learning techniques.

In **Chapter 2** the main ideas behind deep learning are explained, focusing on the specific network architectures and advantages of convolutional neural networks.

**Chapter 3** addresses the problem of detecting individual microcalcifications. The proposed method is based on a combination of a supervised learning technique which was specifically designed to handle efficiently and effectively the computational complexity and the high class imbalance and a supervised deep learning model.

**Chapter 4** still addresses the problem of detecting individual MCs by focusing on the importance of the lesion context. The proposal is a multi-context ensemble of deep neural networks, aiming at learning different levels of the image spatial context, with the goal of improving detection performance.

**Chapter 5** bridges the gap between the detection of individual MCs and the detection and diagnosis of clustered MCs. The proposed approach overcomes the limitations of traditional full CAD scheme, by training an end-to-end system for the detection and classification of MCs clusters.

**Chapter 6** provides a final summary and conclusions.

# Chapter 2

## Deep Learning

---

Deep learning is a growing trend in general data analysis and it is emerging as the leading machine-learning tool in the imaging and computer vision domains. Machine learning is an application of Artificial Intelligence (AI) that provides systems with the ability to automatically learn and improve from experience without being explicitly programmed. Machine-learning technology powers many aspects of modern society and is nowadays involved in many applications. They are used, for example, to identify objects in images, transcribe speech into text and select relevant results of web searches. Conventional machine learning techniques were limited in their ability to process natural data in their raw form. For decades, constructing a pattern recognition or machine learning system required careful engineering and considerable domain expertise to design a feature extractor that transformed the raw data (such as the pixel values of an image) into a suitable internal representation or feature vector from which the learning subsystem, often a classifier, could detect or classify patterns in the input [42]. The attempt of AI researchers has been focused on trying to develop algorithms acting like human intelligence. For example, let us consider a simple task like interpreting a natural image. When humans try to solve such a problem, they usually exploit their intelligence in order to decompose the problem into sub-problems and multiple levels of representation. Humans usually describe a complex concept, task or situation in a hierarchical way, defining several levels of abstraction. It can be assumed that human brain is organized in a deep architecture such that, an input is represented in multiple levels of abstraction, each level corresponding to a different area of the cerebral cortex. Therefore human brain shows to elaborate information coming from a specific situation, through multiple stages of transformation and representation. This is particularly evident when humans manage visualization tasks: the problem is decomposed in several steps, each one detecting more abstract features from edges detection up to complex visual shapes [43]. Therefore a possible and common way to extract useful information from a natural image is to design different modules that transform the raw pixel representation into gradually more abstract representation, starting from the presence of edges, the detection of more complex but local structures, up to the identification of abstract categories associated with objects present in the image. Putting all these representations together allows to build enough understanding of the scene. All these observations lead to the definition of representation learning [44]. Representation learn-

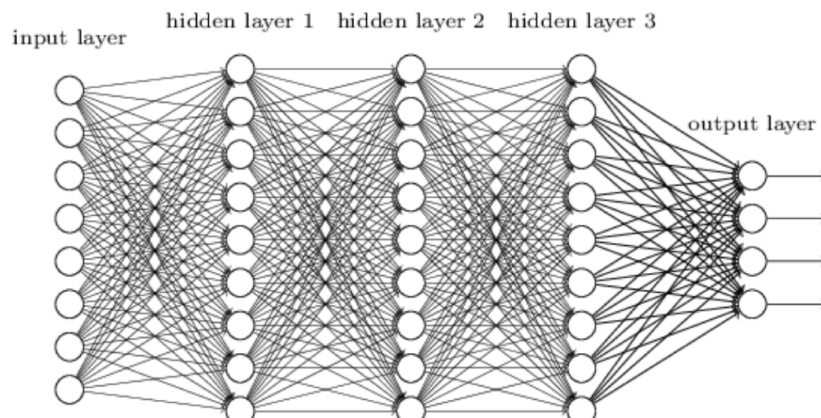


Figure 2.1: An example of deep feed forward neural network

ing consists of a set of methods in which a machine is supposed to take raw data as input and automatically learn the features necessary to perform the detection or the classification. Deep Learning methods are representation learning methods with multiple levels of representation obtained by composing simple but non linear modules such that the complexity of the obtained representation increases with the number of levels designed. The term deep learning identifies computational models that are composed of multiple transformation layers able to learn representations of data with multiple levels of abstraction. Deep learning methods aim at automatically learn high-level features hierarchies by combining lower level features. Automatically learning features at multiple levels of abstraction allows a system to learn very complex functions that map the input to the output directly from data itself, without using human hand-crafted features [42]. Deep learning models are based on deep feedforward neural networks. In the following sections a general description of feedforward neural networks is given, for then focusing on convolutional neural networks, a particular kind of neural networks specifically designed to work with images.

## 2.1 Deep Feedforward Neural Networks

The aim of a feed forward networks is to approximate some function  $y = f(x)$  that, in the case of a classifier, maps an input  $x$  to a category  $y$ . A feedforward network determines a mapping  $y = f(x; \theta)$  and learns the value of the parameters  $\theta$  that results in the best function approximation [45]. These models are called feedforward because there are no feedback connections between the units, which means that the information flows through the function being evaluated from  $x$ , through the intermediate computations used to determine  $f$ , and finally to the output  $y$ . Deep feedforward neural networks are considered the basis of many machine learning applications. This kind of models are mathematically based on the composition of many

different functions, building the final  $f(x)$ , according to a chain structure. Let us suppose to combine three functions  $f^{(1)}$ ,  $f^{(2)}$ ,  $f^{(3)}$  in a chain structure, to form  $f(x) = f^{(3)}(f^{(2)}(f^{(1)}(x)))$ . In this model,  $f^{(1)}$  is defined as the first layer of the network,  $f^{(2)}$  is the second layer, and so on. The overall length chain represents the depth of the model and in general the terminology “deep learning” derives from this structure. The final layer of a feed forward neural network is called the output layer. During the training of the neural network the evaluated function  $f(x)$  is supposed to match  $f^*(x)$ . Each sample  $x$  provides a label  $y \approx f^*(x)$ . The training examples specify directly what the output layer must do at each point  $x$ : it must produce an output value that is as closest as possible to  $y$ . However, the training data does not show the desired output for each layer interposed between the input layer and the output layer. For this reason these intermediate layer are called hidden layers (see Fig. [2.1](#)). Finally, these networks are defined neural because they are inspired to brain neural networks. Each hidden layer is made up of many units, acting in parallel, which are supposed to resemble brain neurons since they receive input from many other units and calculate its own activation value [\[45\]](#). Now it is important to understand the importance of the activation functions that are used to compute the hidden layers values. In modern neural networks the recommendation is to use the Rectified Linear Unit (ReLU) [\[46\]](#), defined by the function

$$g(z) = \max \{0, z\} \tag{2.1}$$

depicted in Fig. [2.2](#). Applying this function to the output of a linear transformation gives a non-linear transformation. The function remains very similar to a linear function, with the main difference that a rectified linear unit outputs zero across half of its domain. This makes ReLU preserving many of the properties that make linear models easy to optimize with gradient-based methods that will be discussed later. Typically, Rectified Linear Units are used on top of an affine transformation in this way:

$$f = g(W^T x + b) \tag{2.2}$$

The learning process of a deep feedforward neural network learns the weights that express the importance of the respective inputs to the output. The aim is to develop a learning algorithm able to find weights and biases so that the output from the network approximates the actual value for all training inputs to be classified.

### 2.1.1 Gradient-based Learning: Stochastic Gradient Descent

The training procedure of a deep feedforward neural network consists of an iterative propagation of samples through the network and modification of its weights, which are properly initialized [\[47\]](#). Deep neural networks are trained using the back-propagation algorithm by minimizing a given objective function, cost function or loss function with respect to the weights  $w$ . For a dataset  $D$ , the optimization objective is the average loss over all  $|D|$  data instances:

$$L(w) = \frac{1}{|D|} \sum_{i=1}^{|D|} f_w(x(i)) \tag{2.3}$$

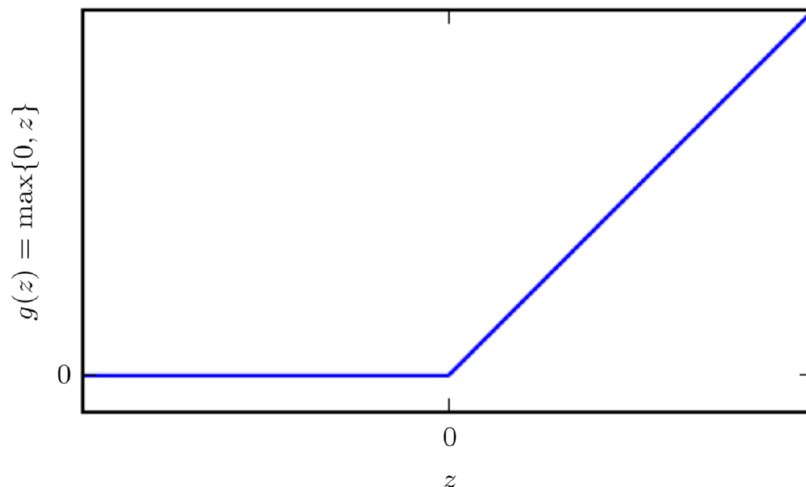


Figure 2.2: ReLU activation function

Since  $D$  can be very large, a stochastic approximation of this objective is used, where the cost over the entire training set is approximated with the cost over mini-batches of data. Drawing a mini-batch of  $N \ll |D|$  instances the optimization function becomes:

$$L(w) \approx \frac{1}{|N|} \sum_{i=1}^{|N|} f_w(x(i)) \quad (2.4)$$

The gradient of a function generalizes the notion of derivative to the case of functions with multiple inputs. Local and global minima of the loss function can be found by moving in the direction of the negative gradient. This is known as the gradient descent method. Training a neural network consists in training the model with gradient descent. Nearly all of deep learning architectures, and also deep feedforward neural networks, are trained using an extension of the gradient descent algorithm: the Stochastic gradient descent (SGD).

The stochastic gradient descent updates the weights  $w$  by a linear combination of the negative gradient  $\nabla L_w$  and the previous weight update  $V_t$  according to the following formula:

$$V_{t+1} = \mu V_t - \alpha \nabla L(w_t) \quad (2.5)$$

where  $\alpha$  and  $\mu$  are two hyperparameters that are chosen for the learning procedure.

## 2.1.2 Learning rate

The coefficient  $\alpha$  is called the learning rate and controls the size of the weight updates (see Fig. [2.3](#)) A too high learning rate will make the learning jump over minima, but a too low learning rate will either take too



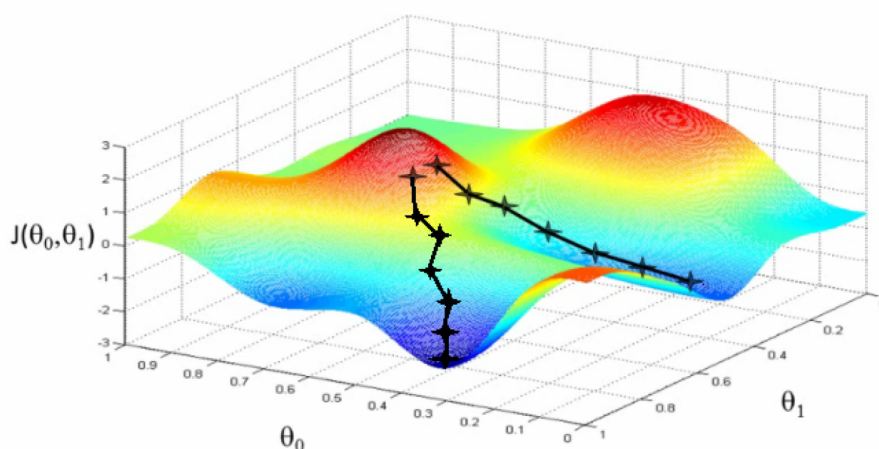


Figure 2.3: Stochastic Gradient descent: the role of learning rate

long to converge or get stuck in an undesirable local minima. In order to achieve faster convergence, prevent oscillations and getting stuck in undesirable local minima the learning rate is often varied during training either in accordance to a learning rate schedule or by using an adaptive learning rate. Common learning rate schedules include time-based decay, step decay and exponential decay.

### 2.1.3 Momentum

The parameter  $\mu$  is the momentum that indicates the contribution of the previous weight update in the current iteration. The momentum algorithm accumulates an exponentially decaying moving average of past gradient and continues to move in their direction as shown in Fig. 2.4 by determining how quickly the contributions of previous gradients exponentially decay [45].

### 2.1.4 Dropout

When training a network with a large number of parameters, an effective regularization mechanism is essential to prevent overfitting. Regularization consists in adding a penalty on the different parameters of the model to reduce the freedom of the model itself and hence reducing probability of fitting the data noise. Classical regularizers such as L1 or L2 regularization have been found to be insufficient in this context. Dropout is a powerful regularization method [48] which has been shown to improve generalization for large neural nets. With dropout, a subset  $p$  of network units is drawn at random and temporarily “switched off” during training [2.5]. When in this state, those units do not propagate signals when a sample is presented, nor participate in the process of error backpropagation. As a result, only a random subset of neurons are trained in a single iteration of SGD by forcing the neural network to learn more robust features that are useful

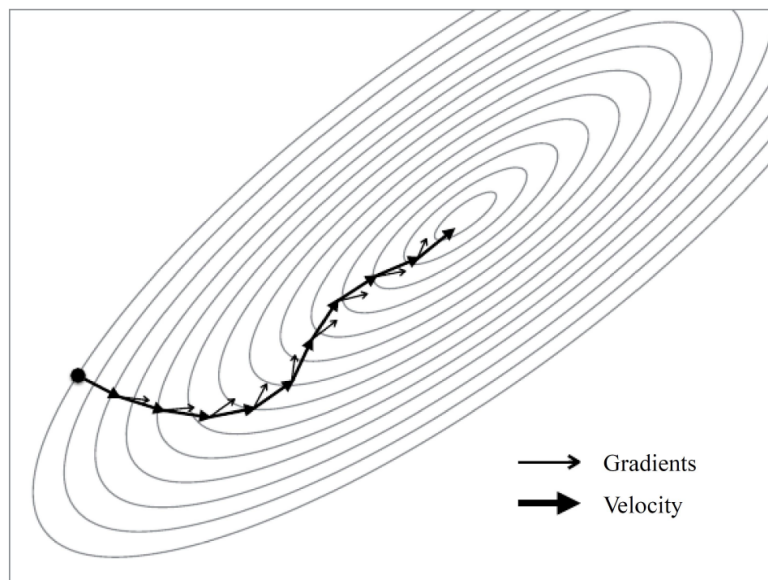


Figure 2.4: Stochastic Gradient descent: the role of momentum

in conjunction with many different random subsets of the other neurons. At test time, all neurons are used, and the activation of each neuron is multiplied by  $p$  to account for the scaling.

## 2.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) [49] are a particular kind of deep neural networks well suited to work with images as they directly take in input 2D or 3D structures, preserving configuration information of the data. CNNs are based on three main architectural ideas: local receptive fields, weight sharing, and subsampling in the spatial domain. A typical CNN principally consists of three types of layers: (i) convolutional layers, (ii) sub-sampling layers, and (iii) output layers, that are arranged in a feed-forward structure [42] (see Fig. 2.6).

### 2.2.1 Convolutional layers

Convolutional layers are responsible for detecting local features in all locations of the input images. To detect local structures, each node in a convolutional layer is connected to only a small subset of spatially connected neurons in the input image channels, called receptive field. Furthermore, to enable the search for the same local feature, connection weights are shared between all the nodes in the convolutional layers; each set of shared weights is called convolutional kernel. For each convolutional layer, a set of convolutional kernels  $W = \{W_1, W_2, \dots, W_n\}$  is convolved with the input image  $X$ , and biases  $B = \{b_1, b_2, \dots, b_n\}$  are added, so as to generate a new feature

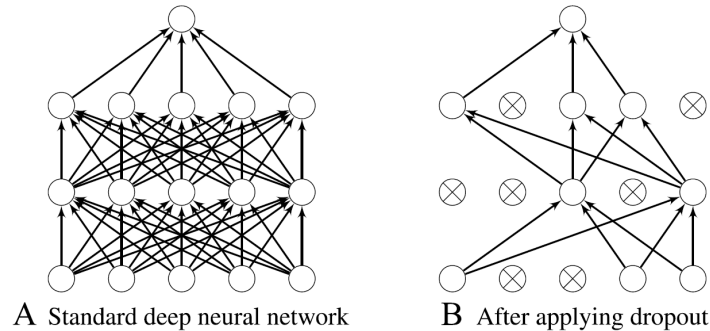


Figure 2.5: Comparison between a standard deep neural network and the same network with dropout application. The circles with a cross symbol inside denote deactivated units.

map  $X_i$  through an element-wise non-linear transform  $\sigma$ :

$$X_i = \sigma(W_i * X + b_i) \quad \forall i = 1, \dots, n \quad (2.6)$$

Due to the local connectivity and weight sharing, the number of parameters compared to a fully connected neural network are greatly reduced, and thus it is possible to avoid overfitting. Further, when the input image is shifted, the activation of the units in the feature maps is also shifted by the same amount, which allows a CNN to be equivariant to small shifts, as illustrated in Fig. 2.7. In the figure, when the pixel values in the input image are shifted by one-pixel right and one-pixel down, the outputs after convolution are also shifted by one-pixel right and one-pixel down.

### 2.2.2 Max pooling layers

Each sequence of convolutional layers is followed by max pooling layers, that are applied to reduce the size of feature maps by selecting the maximum value in local neighbourhoods. Specifically, each feature map in a pooling layer is linked with a feature map in the convolution layer, and each unit in a feature map of the pooling layer is computed based on a subset of units in its receptive field. Similar to the convolution layer, the receptive field that finds a maximal value among the units in its receptive field is convolved with the convolution map but with a stride of the size of the receptive field so that the contiguous receptive fields are not overlapped. The role of the pooling layer is to progressively reduce the spatial size of the feature maps, and thus reduce the number of parameters and computation involved in the network. Another important function of the pooling layer is for translation invariance over small spatial shifts in the input. In Fig. 2.7, while the bottom leftmost image is a translated version of the top leftmost image by one-pixel right and one-pixel down, their outputs after convolution and pooling operations are the same (see units in green).

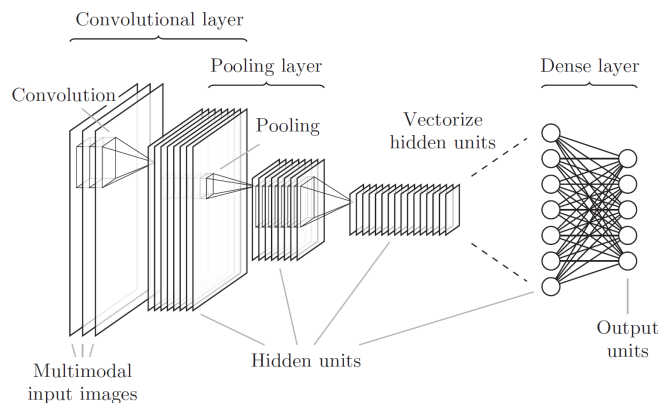


Figure 2.6: Convolutional neural network with two convolutional layers, one pooling layer and one dense layer. The activations of the last layer are the output of the network.

Table 2.1: A list of commonly applied last layer activation functions for various tasks

Task	Last layer activation function
Binary classification	Sigmoid
Multi-class classification	Softmax
Regression to continuous values	Identity

### 2.2.3 Fully connected layers

At the end of the convolutional stream of the network, a number of consecutive fully connected layers is added, and the class distribution over the classes is generated by feeding them through an activation function. Neurons in a fully connected layer have connections to all activations in the previous layer, as in regular non-convolutional artificial neural networks [2.1](#). Their activations can thus be computed as an affine transformation, with matrix multiplication followed by a bias offset.

#### Last layer activation function

The activation function applied to the last fully connected layer is usually different from the others. An appropriate activation function needs to be selected according to each task. An activation function applied to the multiclass classification task is a softmax function which normalizes outputs of the last fully connected layer to target class probabilities, where each value ranges between 0 and 1 and all values sum to 1. Typical choices of the last layer activation function for various types of tasks are summarized in Table [2.1](#).

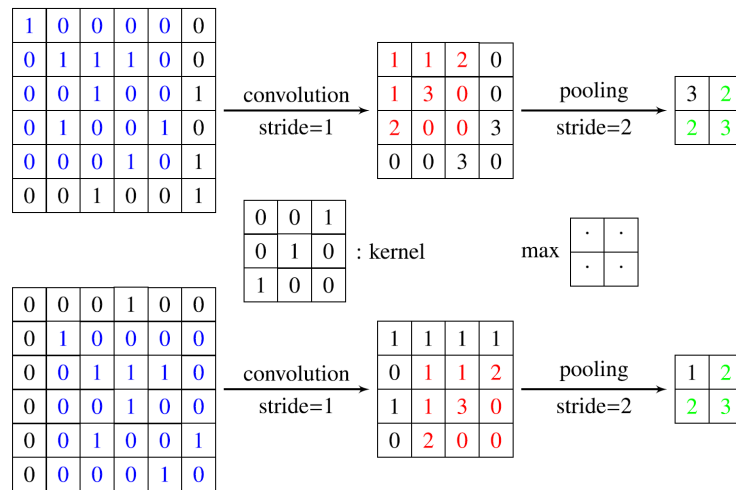


Figure 2.7: Illustration of translation invariance in convolutional neural network. The bottom leftmost input is a translated version of the upper leftmost input image by one-pixel right and one-pixel down.

## 2.3 Deep CNN architectures

When architecting a CNN for a particular task there are multiple factors to consider, including understanding the task to be solved and the requirements to be met and optimize computation and memory footprint. In the early days of modern deep learning it was common to use very simple combinations of the building blocks. Later on, network architectures became much more complex, resulting in updates to the state-of-the-art. In this section a general overview about the best-known CNN network architectures is given, with a particular focus on the ones that are commonly used for medical image tasks, i.e. medical image classification and segmentation.

### 2.3.1 Classification architectures

LeNet [49] and AlexNet [50], introduced over a decade ago, were the very first convolutional architecture to be proposed, being in essence very similar models. Both networks were relatively shallow, consisting of two and five convolutional layers, respectively, and employed kernels with large receptive fields in layers close to the input and smaller kernels closer to the output. AlexNet did incorporate rectified linear units instead of the hyperbolic tangent as activation function, which are now the most common choice in CNNs. After 2012 the exploration of novel architectures took off, and in the last years there is a preference for far deeper models. By stacking smaller kernels, instead of using a single layer of kernels with a large receptive field, a similar function can be represented with less parameters. This approach ensures the same effective receptive field, by increasing at the same time the number of non-linearities (which makes the decision function more discriminative) and decreasing the number of parameters. Simonyan et al. [51] were among the first to explore much deeper networks, and em-

ployed small, fixed size kernels in each layer. This is the idea behind the VGGNet that won the ImageNet challenge of 2014. Rather than using relatively large receptive fields in the first convolutional layers (e.g.  $11 \times 11$  with stride 4 as in Alexnet) they use very small  $3 \times 3$  receptive field throughout the whole net, which are convolved with the input at every pixel, with stride 1 (see Fig. 2.8). It is easy to see that a stack of two  $3 \times 3$  layers (without spatial pooling in between) has an effective receptive field of  $5 \times 5$ ; three such layers have a  $7 \times 7$  effective receptive field, like in the aforementioned Alexnet. Because there are now three ReLU units instead of just one, the decision function is more discriminative. There are also fewer parameters (27 times the number of channels instead of AlexNets 49 times the number of channels). This can be seen as imposing a regularisation on the  $7 \times 7$  convolutional filters, forcing them to have a decomposition through the  $3 \times 3$  (with non-linearity injected in between). On top of the deeper networks, more complex building blocks have been introduced that improve the efficiency of the training procedure and again reduce the amount of parameters. Szegedy et al. [52] introduced a 22-layer network named GoogLeNet, also referred to as Inception, which made use of so-called inception blocks [53], a module that replaces the mapping defined in Eq. 2.6 with a set of convolutions of different sizes. Similar to the stacking of small kernels, this allows a similar function to be represented with less parameters. However as CNNs became increasingly deep, a new research problem emerged: as information about the input or gradient passes through many layers, it can vanish and “wash out” by the time it reaches the end (or beginning) of the network. He et al. [54] addressed the vanishing gradient problem by presenting the ResNet architecture. Resnet introduces skip connections, which makes it possible to train much deeper networks. Having skip connections in addition to the standard pathway gives the network the option to simply copy the activations from layer to layer (more precisely, from ResNet block to ResNet block), preserving information as data goes through the layers. Some features are best constructed in shallow networks, while others require more depth. The skip connections facilitate both at the same time, increasing the network’s flexibility when fed input data. DenseNet was presented in [55] as a logical extension of ResNet. In DenseNet, each layer obtains additional inputs from all preceding layers and passes on its own feature-maps to all subsequent layers. Differently from Resnet that adds the activations produced by one layer to later ones, here concatenation is used. This encourages feature reuse and lowers the number of parameters for a given depth. Xie et al. [56] proposed ResNext, that is an extension of the deep residual network which replaces the standard residual block with one that leverages a “split-transform-merge” strategy used in the Inception models. Squeeze-and-Excitation Networks [57] which won the ILSVRC 2017 competition, builds on ResNext but adds trainable parameters that the network can use to weigh each feature map, where earlier networks simply added them up. These SE-blocks allows the network to model the channel and spatial information separately, increasing the model capacity. SE-blocks can easily be added to any CNN model, with negligible increase in computational costs.

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input ( $224 \times 224$ RGB image)					
conv3-64	conv3-64 <b>LRN</b>	conv3-64 <b>conv3-64</b>	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 <b>conv3-128</b>	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 <b>conv1-256</b>	conv3-256 conv3-256 <b>conv3-256</b>	conv3-256 conv3-256 conv3-256 <b>conv3-256</b>
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 <b>conv1-512</b>	conv3-512 conv3-512 <b>conv3-512</b>	conv3-512 conv3-512 conv3-512 <b>conv3-512</b>
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 <b>conv1-512</b>	conv3-512 conv3-512 <b>conv3-512</b>	conv3-512 conv3-512 conv3-512 <b>conv3-512</b>
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Figure 2.8: VGGnet configurations. The depth of the configurations increases from the left (A) to the right (E), as more layers are added. Source: Simonyan et al. [48]

### 2.3.2 Segmentation architectures

Segmentation is a common task in both natural and medical image analysis and can be defined as a problem of structured prediction where every pixel in the image grid needs to be assigned to a class label. Medical image segmentation, identifying the pixels of organs or lesions from background, is one fundamental step in medical image analysis since it delivers critical information about the shapes and volumes of these lesions, helping clinicians in detecting and diagnosing certain diseases. Segmentation of medical images is a very challenging task due to the large shape and size variations of anatomy between patients. Furthermore, low contrast to surrounding tissues can make automated segmentation even more difficult. To tackle this, CNNs were firstly used to classify each pixel in the image individually, by presenting it with patches extracted around the particular pixel. A drawback of this “sliding-window” approach is that input patches from neighbouring pixels have huge overlap and the same convolutions are computed many times. Fortunately, the convolution and dot product are both



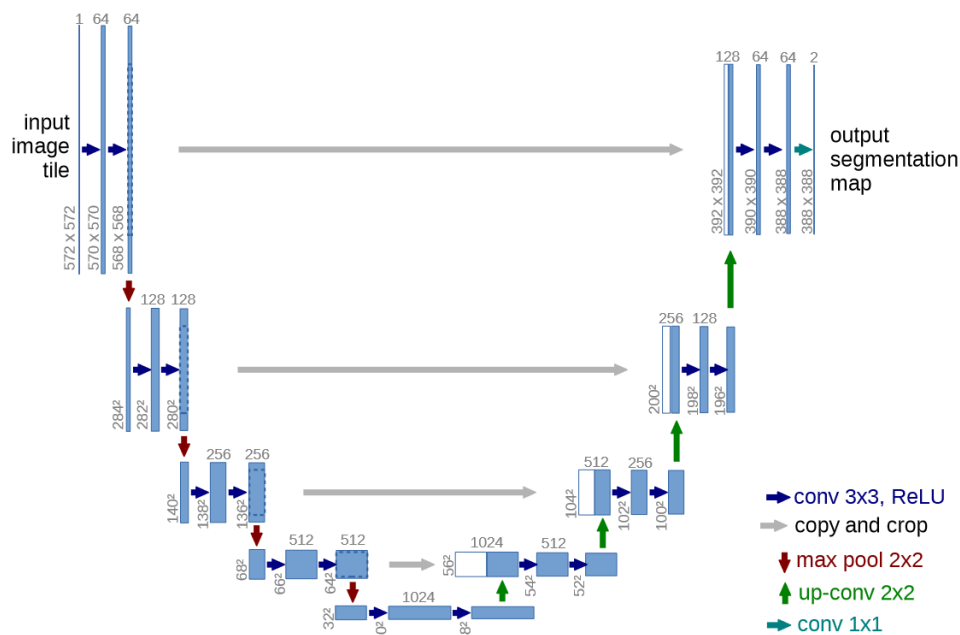


Figure 2.9: U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations. Source: Ronneberger et al. [56]

linear operators and thus inner products can be written as convolutions and vice versa. By rewriting the fully connected layers as convolutions, the CNN can produce likelihood map, rather than an output for a single pixel. The resulting “fully convolutional network” (FCN) can then be applied to an entire input image or volume in an efficient fashion. However, because of pooling layers, this may result in output with a far lower resolution than the input. “Shift-and-stitch” [58] is one of several methods proposed to prevent this decrease in resolution. The FCN is applied to shifted versions of the input image. By stitching the result together, it is possible to obtain a full resolution version of the final output, minus the pixels lost due to the convolutions. Ronneberger et al. [59] took the idea of the FCN one step further and proposed the U-net architecture (see Fig. 2.9). The core of a U-Net is a U-shaped architecture, consisting of a contracting path on the left side (encoder) and of a symmetric expansive path on the right one (decoder). The encoder path consists of the repeated application of double convolution blocks, each one followed by a max pooling operation for downsampling with a pooling size of  $2 \times 2$  and stride of 2. Each double convolution block is made up of two  $3 \times 3$  convolutional kernels followed by a ReLu activation function. This sequence is repeated four times, and after each downsampling filters in the convolutional layers are doubled. The expansive path applies the same blocks but with up-convolution and up-sampling layers. Similar to the encoder, the succession of up-sampling and two convolutional operations is repeated four times but halving the number of filters at each stage. Finally, a  $1 \times 1$  convolution operation is performed for providing the final



segmentation map. The main innovation behind U-Net is however the introduction of shortcut connections. In all the four levels, the output of the convolutional layer prior to the pooling operation of the encoder, is transferred to the decoder and concatenated with the output of the upsampling operation for then propagating the concatenated feature maps to the successive layers. In such a way the network is able to combine deep, semantic, feature maps from the decoder sub-network with shallow, low-level feature maps from the encoder one, resulting to be very effective in recovering fine-grained details of the objects and to generate segmentation masks with fine details even on complex background. A similar approach was used by [60] for 3D data. Milletari et al. [61] proposed an extension to the U-Net layout that incorporates ResNet-like residual blocks and a dice loss layer, rather than the conventional cross-entropy, that directly minimizes this commonly used segmentation error measure.



## Chapter 3

# Computer aided detection of individual microcalcifications

---

Although in the last years many advances have been made in the area of CAD for digital mammograms, the main challenge of accurately identifying individual calcifications still remains and it can be attributed to two main issues. Firstly the appearance of MCs themselves, that appears on mammograms as small bright spots, varying in shape and size, within an inhomogeneous background. Secondly, along with the low prevalence of abnormalities, MCs detection is a severely unbalanced classification problem that depends upon the fact that the positive class is about four orders of magnitude smaller than the negative class, i.e. other breast tissues. This is because the non-MC class includes a large variety of benign calcifications (popcorn-like, rod-like, vascular etc.), different background tissues (fat, dense, connective, etc.) and artifacts (biomarkers, metal clips, dust, etc.). This class skew is a huge problem for most classification strategies that, when trained on highly unbalanced data sets, tend to be overwhelmed by the majority class and have poor performance on the minority class. Several approaches presented for MC detection in the last decades address the class imbalance problem by random undersampling the negative class to obtain approximately the same size for the two classes [62, 63, 64, 65, 66]. Nonetheless, it remains uncertain whether the selected subset is fully representative of the negative class. A different solution was proposed in [67], where a Support Vector Machine (SVM) is repeatedly trained on successively more difficult examples randomly sampled from the negative class. However, the resulting SVM classifier is characterized by a very large number of support vectors, thus making the detection phase computationally intense. Recently, a method specifically designed to effectively learn from heavily unbalanced data was presented in [68]. It consists of a cascade of boosting classifiers (in the following referred to as *Cascade*) with increasing complexity and specificity like in the face detector proposed in [69]. An extension of this approach was presented in [70], where a *Cascade* with a higher number of classifiers (in the following referred to as *Deep Cascade*) achieved state-of-the-art MC detection performance, outperforming widely used high-end CAD commercial systems. This system was used as starting point in this thesis to develop a novel method, presented in [71], that combines the benefits of deep cascade and convolutional neural networks in a novel two-stage classification system, with the aims of reducing the number

## CHAPTER 3. Computer aided detection of individual microcalcifications

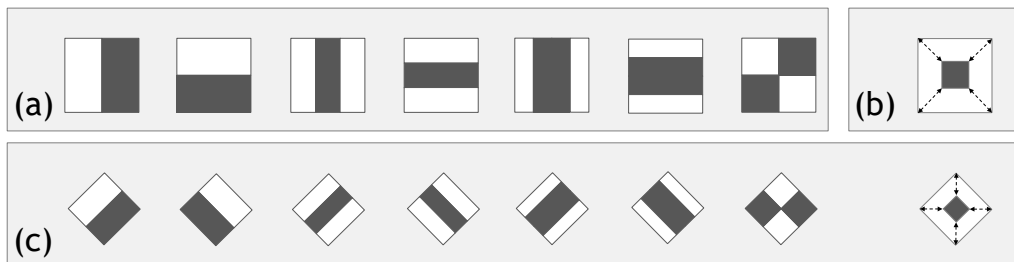


Figure 3.1: The Haar-like feature groups used by the cascade of classifiers. (a) Some examples of the first group. (b) An example of the second group. (c) Some examples of the third group.

of FPs and achieving at the same time a high sensitivity value which can be comparable to human level performance.

### 3.1 The Cascade approach

#### 3.1.1 Ranking based cascade and feature set

The Cascade approach proposed by [68] relies on a two-class ranking-based cascade classifier which classifies each pixel of the mammogram as positive or negative using a subwindow of size  $M \times M$  centered on it, in the following referred to as sample and denoted by  $\mathbf{x}$ . Three groups of Haar-like feature are used. For the first group the value of each feature is calculated as the difference between the sum of pixels belonging to adjacent rectangular regions, aimed at capturing edge and elongated patterns (see Fig. 3.1a). For the second group, the value is calculated in a similar way, but the support regions are two concentric rectangles, so being more suitable for the granule-like shape of MCs (see Fig. 3.1b). The third group is constituted by the 45-rotated version of the feature of the first two groups (see Fig. 3.1c). All features are scaled and translated separately across all possible combinations on the subwindow, obtaining tens of thousands of features. We denote with  $\mathcal{F}$  the set containing such feature. The task of selecting the best feature from  $\mathcal{F}$  to be used in the detection phase is part of the learning phase in each node classifier.

#### 3.1.2 Detection phase

The underlying idea behind cascade is to employ a sequence of node classifiers with different discriminative power  $\{H_i(\mathbf{x})\}_{i=1,\dots,n}$ . A given patch sample  $\mathbf{x}$  passes to the next classifier only if the current one classifies it as MC according to a specific threshold  $\Theta_i$ , that is determined during training so as to achieve a high sensitivity  $s_i$  (usually,  $s_i \geq 99\%$ ). In this way, the most likely-MC samples go through the entire cascade, whereas the easily detectable background patches are discarded by the early stages. As a result, the detection rate  $D$  and false positive rate  $F$  of a cascade composed by  $n$  nodes is given by

$$D = \prod_{i=1}^n d_i \quad F = \prod_{i=1}^n f_i \quad (3.1)$$

In this way, the first nodes of the cascade have to face a simpler task (rejecting the most distinguishable background tissue regions), while the last ones are specialized to discriminate between actual MCs and the background tissue configurations most resembling a MC. This should reduce the number of false positive produced by the detector and concentrate the computational complexity of the system on the last classifier of the cascade.

#### 3.1.3 Learning procedure

Internally, each strong classifier  $H_i(\mathbf{x})$  is a linear combination of  $T_i$  ‘weak’ classifiers  $h_{i,j}(\mathbf{x}) \in \{0, 1\}$  (0 for background tissue, 1 for MC) weighted by  $\alpha_{i,j} \in \mathbb{R}$ :

$$H_i(\mathbf{x}) = \sum_{j=1}^{T_i} \alpha_{i,j} h_{i,j}(\mathbf{x}). \quad (3.2)$$

with  $T_i$  determined during training to keep the false positive rate  $f_i$  below an acceptable level (usually,  $f_i \leq 30\%$ ). The weak classifiers are added in subsequent rounds and are selected during the training time in order to maximize the area under the ROC curve. Such an objective function that is equivalent to the probability of correct pairwise ranking, is insensitive to the class skew, and thus it is a good choice when learning from unbalanced datasets. Moreover, in order to give more value to crucial pairs that are misclassified in the previous rounds a weight distribution is maintained. Each training round the weak classifier that minimizes the weighted sum of misranked crucial pairs on the training set is selected. The number of training rounds for the  $i$  th node is determined by whether the node learning goals  $d$  (detection rate) and  $f$  (false positive rate) have been met. In fact, the decision threshold  $\Theta_i$  of the  $i$ th node classifier  $H_i(\mathbf{x})$  is always chosen at each round so as to meet  $d$ , hence the number of rounds is governed only by whether the condition  $f_{i,t} \leq f$  is satisfied, being  $f_{i,t}$  the actual false positive rate of the  $i$ th node achieved at round  $t$ . However, when the classification task is getting more and more complex throughout the cascade, it could be too difficult to satisfy such a condition, thus causing several unnecessary feature to be added without substantially reducing  $f_{i,t}$ . To solve this problem, new features are added until a significant reduction of  $f_{i,t}$  is achieved. Specifically, let  $\psi_i(\Delta) = \{f_{i,\tau}\}_{\tau=t-\Delta, t-\Delta+1, \dots, t}$  be the latest  $\Delta$  achieved false positive rates of the  $i$ th node we can evaluate the variance  $\sigma_\psi^2 = Var(\psi_i(\Delta))$  and define an early stopping mechanism  $c_{stop}$  if such variance is lower than a small quantity  $\epsilon$ , i.e.,  $\sigma_\psi^2 \leq \epsilon$ . New features are therefore added to the node classifier of the Cascade detector until the condition  $f_{i,t} > f \wedge c_{stop} = \sigma_\psi^2 > \epsilon$  holds.

#### 3.1.4 Deep Cascade

Deep cascade [70] is a cascade where the learning algorithm of each binary pixel classifier has been redesigned in the early stopping mechanism used to avoid overfitting. In this way, it is possible to obtain a large increase of the number of pixel classifiers, and, at the same time, keep unchanged the computational benefit of the cascade and maintain a very low processing time per image. The new system is named deep cascade, where the term deep indicates the high number of classifiers employed in each stage of the

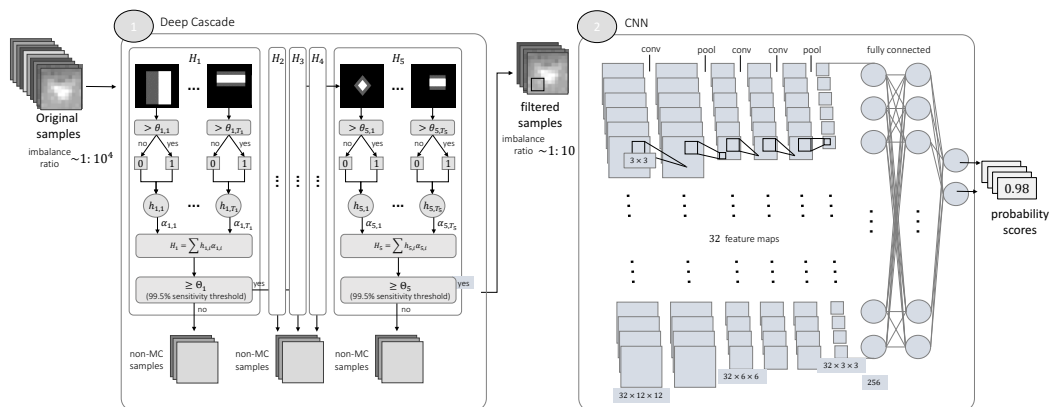


Figure 3.2: Overview of the proposed MC detection scheme. On the left, Deep Cascade reduces the class imbalance ratio in the input data by about three orders of magnitude. This is achieved thanks to a sequence of five high-sensitivity classifiers that linearly combine a large number of decision stumps constrained to use single Haar-like features (top row in each classifier’s box). The remaining samples are then classified by a VGGNet inspired CNN and assigned a probability score using the output from the last fully connected layer.

cascade. In the training phase of the deep cascade detector, the authors redesigned the early stopping mechanism as follows. They removed  $c_{stop}$  and instead established a maximum number of iterations  $T$  to be executed. In summary, new features are added to the node classifier of the deep cascade detector until the condition  $f_{i,t} > f$  or  $t \leq T$ . If the condition  $t = T$  is reached, i.e., the learning goals are not met after  $T$  iterations, it reverts to the training round  $t$  for which  $f_{i,t}$  is minimum.

## 3.2 Combining Deep Cascade and Convolutional Neural Networks

Recently, both deep cascade classifiers and convolutional neural networks have achieved powerful microcalcification detection performance in digital mammography [72, 73]. In this work, we introduce a two-stage classification scheme that combines the benefits of both systems. On one hand, deep cascade showed to be successful in effectively learn from heavily unbalanced data as in the case of MCs ( $\sim 1$  MC every 10,000 non-MC samples). On the other hand, CNNs are powerful models that achieve impressive results for image classification thanks to the ability to automatically extract general-purpose features from the data, but require balanced classes. To overcome these limitations a novel two-stage classification scheme (see Fig. 3.2) is proposed, in which *Deep Cascade* is used to discard most of the negative samples and then train a CNN on almost balanced dataset. Firstly, deep cascades are trained by requiring a very high sensitivity (99.5%) throughout

## 3.2 Combining Deep Cascade and Convolutional Neural Networks

---

the sequence of classifiers. As a result, while the number of MC samples remains practically unchanged, the number of non-MC samples is greatly reduced, reducing the class imbalance ratio by about three orders of magnitude (from  $\sim 1 : 10^4$  to  $\sim 1 : 10$ ). The remaining samples are then used to train a CNN that specializes to distinguish between MCs and the most confusing background tissue patches. The same two-stage classification is also adopted for testing, and is compared to current state-of-the-art MC detection approaches.

In this work a *Deep Cascade* with  $n = 5$ ,  $s_i = 0.995$ , and  $f_i = 0.25$  (see Fig. 3.2) is employed. As a consequence, the overall sensitivity  $S$  and false positive rate  $F$  of the classifier are:

$$S = \prod_{i=1}^5 s_i \quad F = \prod_{i=1}^5 f_i \quad (3.3)$$

which yields  $S = 0.975$  and  $F = 0.001$ . In other words, the number of MC samples remains practically unchanged, while the number of background samples is reduced by three orders of magnitude along with the imbalance ratio which passes from  $\sim 1 : 10^4$  to  $\sim 1 : 10$ .

The output patches filtered by the Deep Cascade are then fed in input to a CNN. In this study, a CNN inspired by the VGGNet architecture [51] is implemented. The CNN model consists of two stacks of two convolutional layers followed by one max-pooling layer. ReLU is used as activation function for each convolutional layer. The final layers are three fully connected layers. The parameters of each layer are shown in Table 3.1.

### 3.2.1 Materials

For this study, 1,066 mammograms acquired with GE Senographe systems (GE, Fairfield, Connecticut, United States) were collected. All available medio-lateral oblique and cranial caudal views of the left and right breast were included. The images were acquired with standard clinical settings in Radboud University Medical Center (Nijmegen, The Netherlands) after referral in screening. Only unprocessed raw FFDM images were used in this study. In all mammograms, a total of 7,579 individual MCs were annotated by experienced readers who marked the center of each MC based on the diagnostic reports.

### 3.2.2 Experiments

For the sake of comparison, the MC detection performance of the proposed two-stage approach (hereafter abbreviated as DC-CNN) is compared with the one of a standalone *Deep Cascade* (hereafter abbreviated as DC) and of a standalone CNN with the same architecture used in the proposed approach.

#### Training and test sets

All methods were trained and tested on patches of size  $12 \times 12$  pixels extracted from the mammograms so that each MC was contained in one patch. All mammograms were preprocessed to remove the noise dependency on the intensity [74]. MC patches were extracted by centering the window on

## CHAPTER 3. Computer aided detection of individual microcalcifications

---

Table 3.1: Architecture of the VGGNet-based CNN

Layer	Type	Output size	Kernel Size	Stride	Padding
0	Input	$1 \times 12 \times 12$			
1	Convolutional	$32 \times 12 \times 12$	$3 \times 3$	1	1
2	ReLU	$32 \times 12 \times 12$			
3	Convolutional	$32 \times 12 \times 12$	$3 \times 3$	1	1
4	ReLU	$32 \times 12 \times 12$			
5	Max pooling	$32 \times 6 \times 6$	$2 \times 2$	2	1
6	Convolutional	$32 \times 6 \times 6$	$3 \times 3$	1	1
7	ReLU	$32 \times 6 \times 6$			
8	Convolutional	$32 \times 6 \times 6$	$3 \times 3$	1	1
9	ReLU	$32 \times 6 \times 6$			
10	Max pooling	$32 \times 3 \times 3$	$2 \times 2$	2	1
11	Fully connected	256	$1 \times 1$		
12	Dropout	256			
13	Fully connected	256	$1 \times 1$		
14	Dropout	256			
15	Fully connected	2	$1 \times 1$		

the annotated MC centers, yielding 7,579 MC patches. Background tissue patches were extracted from the remaining regions of the image with overlapping sliding windows, totalizing 27,017,503 non-MC patches. We applied case-based 2-fold cross validation in all experiments. In each cross validation step, the MC detector was trained on the 50% of the cases, and tested on the other 50%. When splitting the data into a training and test set, the patches belonging to the same case were assigned to the same set.

### Training parameters

DC detectors were trained using sensitivity  $s_i = 0.995$  and false positive rate  $f_i = 0.25$  for each strong classifier. In DC-CNN, the cascade was composed by  $n = 5$  strong classifiers so as to achieve an overall false positive rate  $F = 10^{-3}$ , whereas in DC, according to [70], more classifiers were added to the cascade until all samples from the training set were used, yielding  $n = 6.5$  on average. The total number of decision stumps was on average 2, 530 and 4, 255 for DC-CNN and DC, respectively. The CNNs were trained on perfectly balanced data sets. Augmentation of the positive class was performed by randomly flipping the MC patches horizontally and vertically and by randomly rotating the patches  $90^\circ$ ,  $180^\circ$ , and  $270^\circ$ . The learning algorithm was backpropagation with Stochastic Gradient Descent (SGD) and weight updates that proceeded in mini-batches of 32 patches. In each mini-batch the number of positive and negative samples were balanced. The learning rate was initially set to  $10^{-3}$  and decreased by a factor of 10 every 6 epochs. In total, the learning rate was decreased 5 times, and the learning was stopped after 30 epochs. Momentum and weight decay were set respectively to 0.9 and  $5 \times 10^{-4}$ .



Table 3.2: Comparative results of mean MC detection sensitivity  $\bar{S}$ 

Method	$\bar{S}$	Compared to	Difference	$p$ -value
DC	76.79	-	-	-
CNN	77.53	-	-	-
DC-CNN	<b>79.98</b>	DC	+3.19	< <b>0.001</b>
		CNN	+2.45	< <b>0.001</b>

Table 3.3: Average per-mammogram processing time

Method	Parallelization	Time (s)
DC	-	8.7
DC	CPU (4× Intel(R) Xeon(R) CPU E5-4610 v2)	2.5
CNN	-	2420.9
CNN	GPU (1× NVIDIA TitanX)	5.7
DC-CNN	-	7.2
DC-CNN	CPU/GPU	<b>2.0</b>

### Performance Evaluation

MCs detectors were evaluated in terms of Receiver Operating Characteristics (ROC) curve by plotting True Positive Rate (TPR) against False Positive Rate (FPR) for a series of thresholds on the detector output associated to each sample. Furthermore, the mean sensitivity of the ROC curve in the specificity range on a logarithmic scale was calculated and compared as in [1.1](#). The range  $[a, b]$  was set to  $[10^{-6}, 10^{-1}]$  corresponding to a wide range of operating points that can be used for further analysis by the CAD system. Statistical comparison was performed by means of bootstrapping [\[75\]](#). On the test set, average ROC curves were calculated over 1,000 bootstraps. Additionally, the mean sensitivity was calculated for each bootstrap and  $p$ -values were computed for testing significance. The statistical significance level was chosen as  $\alpha = 0.05$ , but performance differences were considered statistically significant if  $p < 0.025$  due to the Bonferroni correction<sup>1</sup> [\[76\]](#).

## 3.3 Results

The comparative results of mean MC detection sensitivity obtained from the ROC analysis are shown in [Table 3.2](#). The proposed DC-CNN approach achieved a significantly higher mean sensitivity compared to both DC and CNN, yielding an improvement of 3.19% and 2.45%, respectively. ROC curves are shown in [Fig. 4.6](#) and plotted on a logarithmic scale to show the difference between the methods at high specificity. At a false positive rate of

<sup>1</sup>the significance level was obtained as  $\alpha$  divided by the number of comparisons (2 in our case).

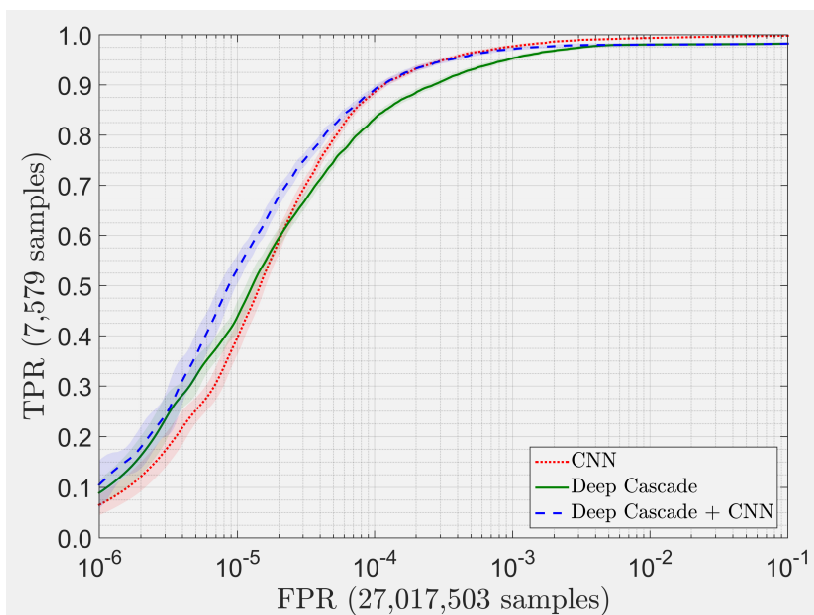


Figure 3.3: Average ROC curves obtained from 1,000 bootstrap iterations. Confidence bands indicate 95% confidence intervals along the TPR axis.

$10^{-4}$  and  $10^{-5}$ , DC-CNN yielded substantial improvements in true positive rate, by 5.75% and 9.49% over DC, and by 0.53% and 13.63% over CNN, respectively. For completeness, in Table 3.3 the average per-mammogram processing times are also reported. Remarkably, DC-CNN was significantly faster than CNN ( $-3.7s$ ) and slightly faster than DC ( $-0.5s$ ). DC-CNN was implemented in C++ using the OpenCV library [77] and the Caffe framework [78].

### 3.4 Discussion

Computer-Aided Detection of MCs in digital mammography can be a valuable tool for radiologists when reading screening mammograms. Usually, CAD systems are made up of two stages: (i) detection of MC candidates; and (ii) classification of MC groups into benign and malignant [68]. In this work, we focused on the detection of MC candidates and improved the state-of-the-art performance by combining two of the latest cutting edge MC detection techniques, namely *Deep Cascade* and CNNs. Remarkably, this result was achieved without generating further computational workload. A slightly shorter *Deep Cascade* is used to discard the majority of non-MC samples, and a CNN to classify the few remaining samples. In this way, the two systems combined were faster than the standalone versions.

It is reasonable to believe that the improvements in MC detection performance obtained in this work could be easily transferred to a full CAD scheme, similarly to what has been done in [70]. In addition, other variants of the proposed DC-CNN scheme could be analysed, e.g. by reducing or increasing the number of strong classifiers in the DC, and by employing other CNN architectures.

## Chapter 4

# Multi-context ensemble of CNNs for improving the automated detection of individual microcalcifications

---

Deep learning models, and in particular CNNs, have recently acquired great popularity thanks to their remarkable performance in computer vision [42, 79] and have proved to be powerful also in medical image analysis [13, 15, 14, 80, 81, 82]. The reason behind this success is the capability of learning hierarchical feature representations directly from data, instead of using handcrafted features based on domain-specific knowledge. As described in Chapter 2, the typical CNN architecture for image processing consists of a series of layers of convolutional filters spaced with downsampling layers. Convolutional filters are applied to small patches of the input images (containing candidate lesion or background) and are able to build features with increasing relevance, from texture to higher order features like local and global shape. The output of the CNN is typically one or more values that represent the probability that an image patch contains a lesion or not.

In this context, patch dimensions play an important role, especially when the lesion is particularly small and similar to the surrounding tissue. If the patch is defined to strictly contain the lesion, it may be too small to produce a set of sufficiently discriminating representations. On the other hand, a larger patch would include much more background which can bias the detection system to focus on uninteresting details contained in the background part. As a consequence, the number of background patches erroneously detected as lesions may be high and limit the benefits that the CADe system can provide, even when deep learning techniques are applied [83, 84]

A simple yet effective way, commonly used in Machine Learning, for boosting the performance of poor detection models is the so called “expert combination”: multiple detectors are trained by using different weight settings and/or different partitions of the same data and strategically combined to solve a particular detection problem [85, 86, 87, 88, 89]. The rationale is that differently trained networks can learn different representations of the training data and, in this way, can agree on correct predictions and make

their errors in different parts of the input space. When combined together, such diversity enforces the correct predictions and reduces the errors, minimizing the risk due to poor model selection. This approach is also useful in medical image analysis field, where ensembles of CNNs have been used to solve many medical image analysis tasks [90, 91, 92].

In this thesis, an approach for the automated detection of MCs in digital mammograms is presented, consisting of an ensemble of CNNs, each one specifically designed to learn a different view of the same lesion. Patches of different dimensions, centered at the same detection location, are extracted to separately train different CNNs, whose network architectures are tailored to the dimensions of the input samples. The idea is that, starting from image patches small enough to entirely contain the lesion to be detected, the size of the neighbourhood is progressively enlarged, and the depth of the network is increased at the same time. In this way, shallower networks become specialized in learning local image features, whereas deeper ones are well suited to learn patterns of the contextual background tissues. Once trained, the detectors are combined together to obtain a final ensemble that can effectively detect abnormalities with a substantial reduction of false positive regions (thanks to the diversity provided by the different spatial context learned by each network). The proposed multi-context ensemble was firstly introduced in [93] and further investigated in [94].

Recently, few other works have tried to add contextual information into the training phase. In [95] a two-pathway CNN architecture for brain tumour segmentation was proposed. Similarly, Kamnitsas et al. [96] employed a dual pathway architecture that processes 3-D input images at multiple scales simultaneously for accurate brain lesion segmentation. Wang et al. [97] proposed a context-sensitive DNN for microcalcification detection by merging, at training time, features coming from two different subnetworks.

The proposed approach stands out from these works since the networks are separately trained and the probability scores are merged at inference time, by allowing to focus on more different portions of the lesion background, without requiring a high computational burden and resulting in a more discriminating power.

The rest of the chapter is organised as follows. Sect. 4.1 introduces the underlying concepts of the proposed method along with a detailed characterization of the proposed architecture. Sect. 4.2 reports the experimental analysis, followed by results in Sect. 5.5. Finally, Sect. 4.4 ends the chapter with discussion and conclusions.

### 4.1 Multi-context CNN ensemble

In this section, the proposed multi-context CNN ensemble for the detection of individual MCs on digital mammograms is presented. Specifically, the proposed ensemble consists of  $K$  different CNNs that are meant to focus on different spatial context of the images and thus to specialize both on local features and on contextual ones. To this end, each network of the ensemble is trained by using image patches of different size, aiming to capture the spatial context around the same detection location. Furthermore, according to the image patch dimensions, the  $K$  network architectures are set to different levels of depth, with the aim of using deeper, hence more discriminating, networks to manage larger image windows.

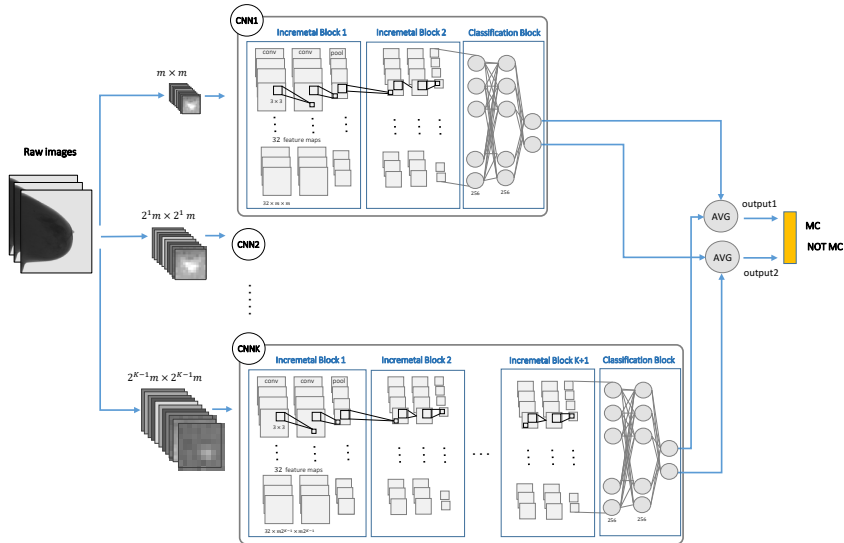


Figure 4.1: Overview of the proposed architecture

The size  $m$  of the smallest patches used in the ensemble is chosen to entirely contain a single lesion, and then it is progressively enlarged to include larger image portions, up to a dimension that is still representative of the context around the lesion. Similarly, the network architecture is set to a baseline configuration, and then its depth is increased as the image size grows. The baseline configuration is inspired by the VGGNet [51], and it is defined as two blocks of two convolutional layers, interlaced by a ReLU activation function and followed by a max pooling layer. Each of these blocks has been named *incremental block*; the word *incremental* indicating they are added to the stack of layers in order to define deeper architectures. More details of the structure of an incremental block are given in Table 4.1. Following the design approach defined by the VGGNet [51], small  $3 \times 3$  kernels are used in each block, since they are faster to convolve with and contain less weights. For the same purpose of decreasing the amount of computations, data reduction layers need to be set to steadily decrease the spatial resolution of the input feature maps. Let  $s_{in}$  be the size of an image patch in input to a convolutional layer or a max pooling layer, we know that its output dimensions can be expressed as:

$$s_{out} = \frac{s_{in} + 2 * pad - kernel}{stride} + 1 \quad (4.1)$$

where  $kernel$  indicates the size of the filter,  $pad$  specifies the padding size, and  $stride$  the intervals at which the filter is applied. We set the stride of convolutional layers equal to 1, by fixing instead the stride of max pooling layers equal to 2 (see Table 4.1). As a result, the image patches are halved after each passage through an incremental block. As a consequence, we decided to progressively double the size of the input patches every time we added an incremental block to the baseline network architecture. To summarize, we can say that the proposed ensemble of CNNs consists of  $K$  different networks, each one trained on image patches of size  $s = \{2^{i-1}m \times 2^{i-1}m\}$  and built with  $d = i + 1$  incremental blocks,

Table 4.1: Details of the *incremental block*

Layer	Type	Output size	Kernel Size	Stride	Padding
1	Convolutional	$32 \times m \times m$	$3 \times 3$	1	1
2	ReLU	$32 \times m \times m$			
3	Convolutional	$32 \times m \times m$	$3 \times 3$	1	1
4	ReLU	$32 \times m \times m$			
5	Max pooling	$32 \times \frac{m}{2} \times \frac{m}{2}$	$2 \times 2$	2	1

Table 4.2: Details of the *classification block*

Layer	Type	Output size	Kernel Size	Rate
1	Fully connected	256	$1 \times 1$	
2	Dropout	256		0.5
3	Fully connected	256	$1 \times 1$	
4	Dropout	256		0.5
5	Fully connected	2	$1 \times 1$	

$\forall i = 1, 2, \dots, K$ .

Each of the  $K$  networks ends with a *classification block*, i.e., with three fully connected layers intertwined with two dropout layers. At the end, a softmax function is applied to the two-output neurons to generate a two-value probability vector associated to each prediction. More details on the classification block are reported in Table 4.2. The  $K$  nets are individually trained and the output values  $Y_i, \forall i = 1, \dots, K$  of the  $K$  CNNs are merged together at inference time to aggregate the multi-level contextual information for the final classification. In particular, the probability values are averaged, resulting in a single probability vector  $Y_{en} = \{Y_{en,p}, Y_{en,n}\}$  associated to each patch, stating the final decision about that sample:

$$Y_{en} = \{Y_{en,p}, Y_{en,n}\} = \left\{ \sum_{i=1}^K \frac{Y_{i,p}}{K}, \sum_{i=1}^K \frac{Y_{i,n}}{K} \right\} \quad (4.2)$$

## 4.2 Experimental analysis

### 4.2.1 Dataset

The publicly available INbreast database [98] was used for this study. The InBreast database was acquired from the Breast Centre of the university hospital of Porto, between April 2008 and July 2010 under permission of both the Hospitals Ethics Committee and the National Committee of Data Protection. The acquisition equipment was the Mammo Novation Siemens FFDM, with a solid-state detector of amorphous selenium. The image matrix was of  $3328 \times 4084$  or  $2560 \times 3328$  pixels, with a pixel-size of  $70 \mu\text{m}$  and a 14-bit contrast resolution, depending on the compression plate used

## 4.2 Experimental analysis

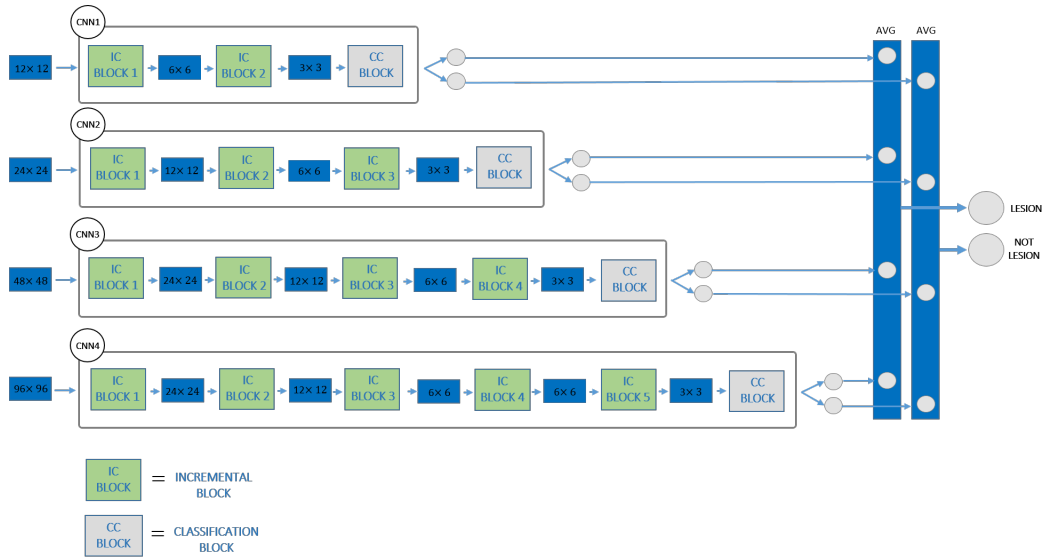


Figure 4.2: Details of the proposed architecture

in the acquisition (according to the breast size of the patient). Images were saved in the DICOM format [4.4](#) and all confidential medical information was removed from the DICOM file. INbreast has FFDM images from screening, diagnostic, and follow-up cases. Screening is made according to national and regional standards. Diagnostic is made when screening shows signs of anomaly. In follow-up images, cancer was previously detected and treated. A total of 115 cases were collected, from which 90 have two images (MLO and CC) of each breast and the remaining 25 cases are from women who had a mastectomy and two views of only one breast were included. This sums to a total of 410 images. Eight of the 91 cases with 2 images per breast also have images acquired in different timings (follow-up). The database includes examples of normal mammograms, mammograms with masses, mammograms with calcifications, architectural distortions, asymmetries, and images with multiple findings. The graphic in Fig. [4.3](#) shows that there is a big prominence of calcifications in the database. This reflects the real population, where calcifications are the most common finding in mammography. Among the 410 images, calcifications can be found in 301 images, and a total of 6,880 individual calcifications have been identified. The main characteristic of the dataset is the carefully associated GT annotation. The annotations were made by a specialist in the field, and validated by a second specialist, between April 2010 and December 2010. When there was a disagreement between the experts, the case was discussed until a consensus was obtained. Each finding has a label that identifies the type of lesion. For the calcification a detailed contour of the finding was made. An ellipse enclosing the entire cluster was also adopted to annotate the clusters of MCCs [4.5](#).

For our experiments, all the images were used and image patches were extracted from the mammograms to train the CNNs. Each patch was labeled as positive or negative according to the information provided by the ground-truth. MC patches were extracted by centering the windows on the annotated MC centers, whereas background tissue patches were extracted from the remaining regions of the images with overlapping sliding windows.



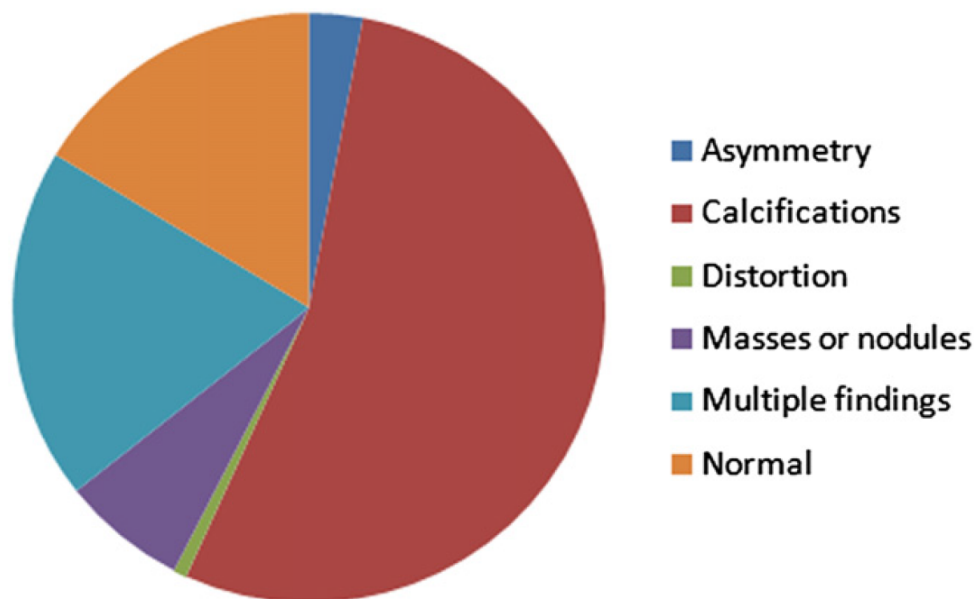


Figure 4.3: Chart describing the findings in the INbreast database

According to the multi-patch criterion, different subwindows of different size were extracted around the same center, by yielding 5,628 positive samples and 26,887,769 negative ones. The resulting patches were used to train and test the proposed detection system.

## 4.2.2 Network architecture

The proposed model consists of a multiple pathway of  $K$  specialized CNNs, each learning a different context extracted from an increasing area centered on the lesion. The choice of the size  $m$  is made in order to guarantee that input patches entirely contain at least the smallest lesions. Considering the size of MCs and the image spatial resolution, we found that a patch size of  $m = 12$  pixels was sufficient to cover the extent of the smallest lesions and to focus on their fine details. Then, the input size is enlarged for the other CNNs to capture larger MCs as well as their background context, by doubling the patch dimensions up to 96 pixels. Larger image portions were not considered since they were not representative of the background context of the lesions and to maintain a reasonable processing time (see Table 4.8). In summary, the patch size ranges from  $12 \times 12$  to  $96 \times 96$ , resulting in a final ensemble made up of  $K = 4$  networks, the first ones more focused on learning details of the lesions and the others on learning background patterns.

The final architecture of the ensemble along with the dimension details of each CNNs are illustrated in Fig. 4.2.



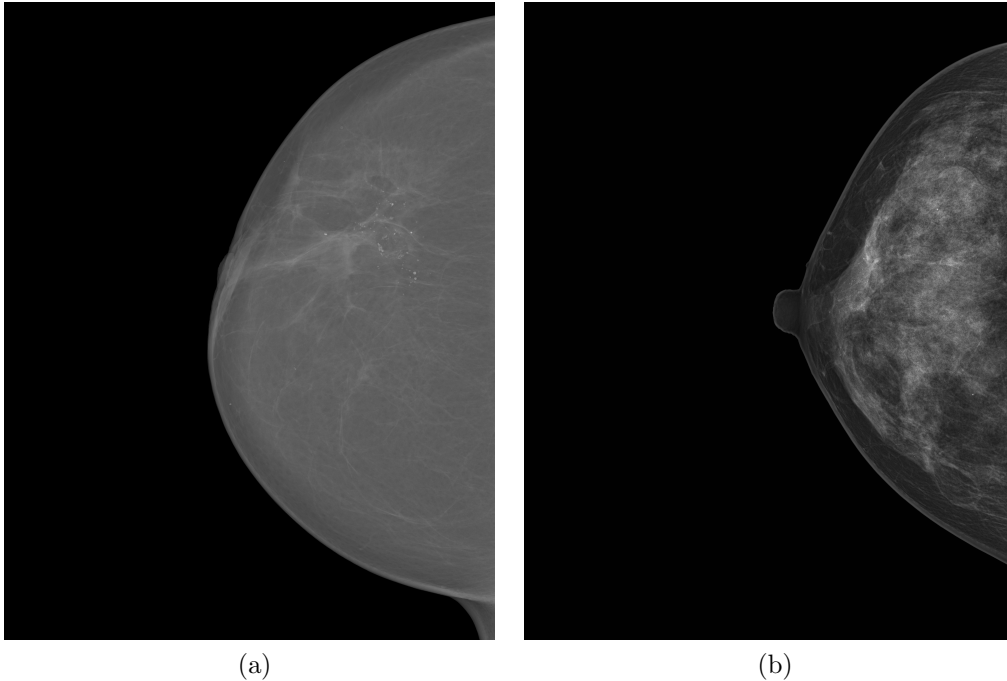


Figure 4.4: Some examples of images from (a-b) INbreast

### 4.2.3 Training parameters

According to the number of extracted patches, MCs detection is a heavily unbalanced classification problems. To avoid the classifiers being overwhelmed by the majority class and misclassify the samples of the minority class, data augmentation was applied, by restoring the balance between positive and negative samples. Thus, all the CNNs of the ensemble were trained on a perfectly balanced dataset. Augmentation of the positive class was performed by randomly flipping the patches horizontally and vertically and by randomly rotating the patches  $90^\circ$ ,  $180^\circ$ , and  $270^\circ$ . Once generated, image patches were standardized by mean subtraction and normalization to unit variance [99].

As to weight initialization and training parameters, all the CNNs of the ensemble were treated in the same way. For all weights in all the layers Xavier initialization [47] was used, while each CNN was optimized to minimize the Softmax loss function by using backpropagation and Mini-Batch Stochastic Gradient Descent. The mini-batch size was of 32 samples and in each mini-batch positive and negative samples were balanced. The learning rate was set to the initial value of  $10^{-3}$  and decreased during training by a factor of 10 every 6 epochs. The learning was stopped after 30 epochs. Momentum and weight decay were set respectively to 0.9 and  $5 \times 10^{-4}$ . The number of feature maps was set to 32, whereas dropout was performed with a probability of 0.5 indicating that, at each training stage, half of the units coming from the previous layer were ignored in the training of the successive layer. The proposed architecture was implemented with a modified version of the Caffe framework [100], and the experiments were conducted on a machine with 2 Intel Xeon e5-2609, 256 GB of RAM and 2 GPU NVIDIA Titan Xp.

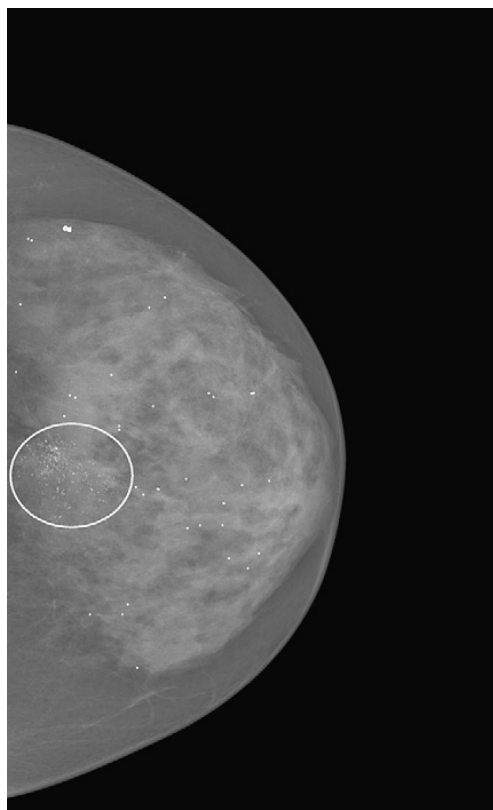


Figure 4.5: InBreast annotation example

### 4.3 Results

To evaluate the performance of the proposed ensemble, an image-based 2-fold cross validation was applied for all the experiments. In each cross validation step, each detector was trained on the 50% of the images and tested on the other 50%. When splitting the data into training and test sets, the patches belonging to the same image were assigned to the same set.

The detectors were evaluated in terms of Receiver Operating Characteristics (ROC) curve and it is worth remarking that the ROC curves were calculated using the image patches. The number of negative and positive patches tested are the same of the original dataset as reported in Table 5.2 (the two leftmost columns). Furthermore, the mean sensitivity of the ROC curve in the specificity range on a logarithmic scale was calculated and compared. The range  $[a, b]$  in eq. 1.1 was set to  $[10^{-6}, 10^{-1}]$  corresponding to a wide range of operating points that are close to practical application requirements of CADe systems for the problem under consideration [101].

For the experimental evaluation, the performances of the standalone CNNs were firstly investigated, by varying the input patch size along with the network depth. In Table 4.3, the performance of the individually trained CNNs for growing values of patch size and network depth are reported. We can see that using larger patches with a deeper network is initially beneficial to improve detection performance. The mean sensitivity increases from 76.30% of CNN1 to 77.45% of CNN3. However, increasing the size of the image window stops to be beneficial and performance decreases. The mean

Table 4.3: Results of mean MC detection sensitivity  $\bar{S}$  for standalone CNNs

Method	patch size	$d$	$\bar{S}_{MC}$
CNN1	$12 \times 12$	2	76.30
CNN2	$24 \times 24$	3	76.90
CNN3	$48 \times 48$	4	77.45
CNN4	$96 \times 96$	5	75.83

Table 4.4: Results of mean MC sensitivity  $\bar{S}$  for combined CNNs

Method	patch size	$d$	$\bar{S}_{MC}$
CNN1+CNN2	12+24	2+3	79.51
CNN1+CNN2+CNN3	12+24+48	2+3+4	81.39
CNN1+CNN2+CNN3+CNN4	12+24+48+96	2+3+4+5	<b>83.54</b>

sensitivity reduced from 77.45% of CNN3 to 75.83% of CNN4.

Furthermore, to understand how joint predictions of the individual pathways affects the performance, in Table 4.4 the results obtained by combining the single CNNs are also reported. We can see that detection performance increases each time a new CNN is added to the ensemble, obtaining the best performance measure when all the networks are used. The proposed full architecture achieved a mean sensitivity of 83.54%. It is worth noting that, even when a single CNN does not perform very well (as in the extreme cases of patch size 12 and 96) they still give a contribution when added to the ensemble.

For the sake of completeness, the effect on the proposed approach of different combination methods in addition to the mean rule is also investigated. In particular, the probability values of the standalone CNNs were combined with the following rules [85]: (i) trimmed mean; (ii) maximum; (iii) minimum; and (iv) majority voting. Results are reported in Table 4.5 showing that the mean rule gave the best performance.

To evaluate the performance of the proposed approach with respect to the literature, we compared our ensemble method with the deep network

Table 4.5: Results of MC sensitivity  $\bar{S}$  for combined CNNs according to different combination rules

CNN1+CNN2+CNN3+CNN4	$\bar{S}_{MC}$
mean	<b>83.54</b>
trimmed mean	81.92
max	77.51
min	81.27
majority voting	81.25

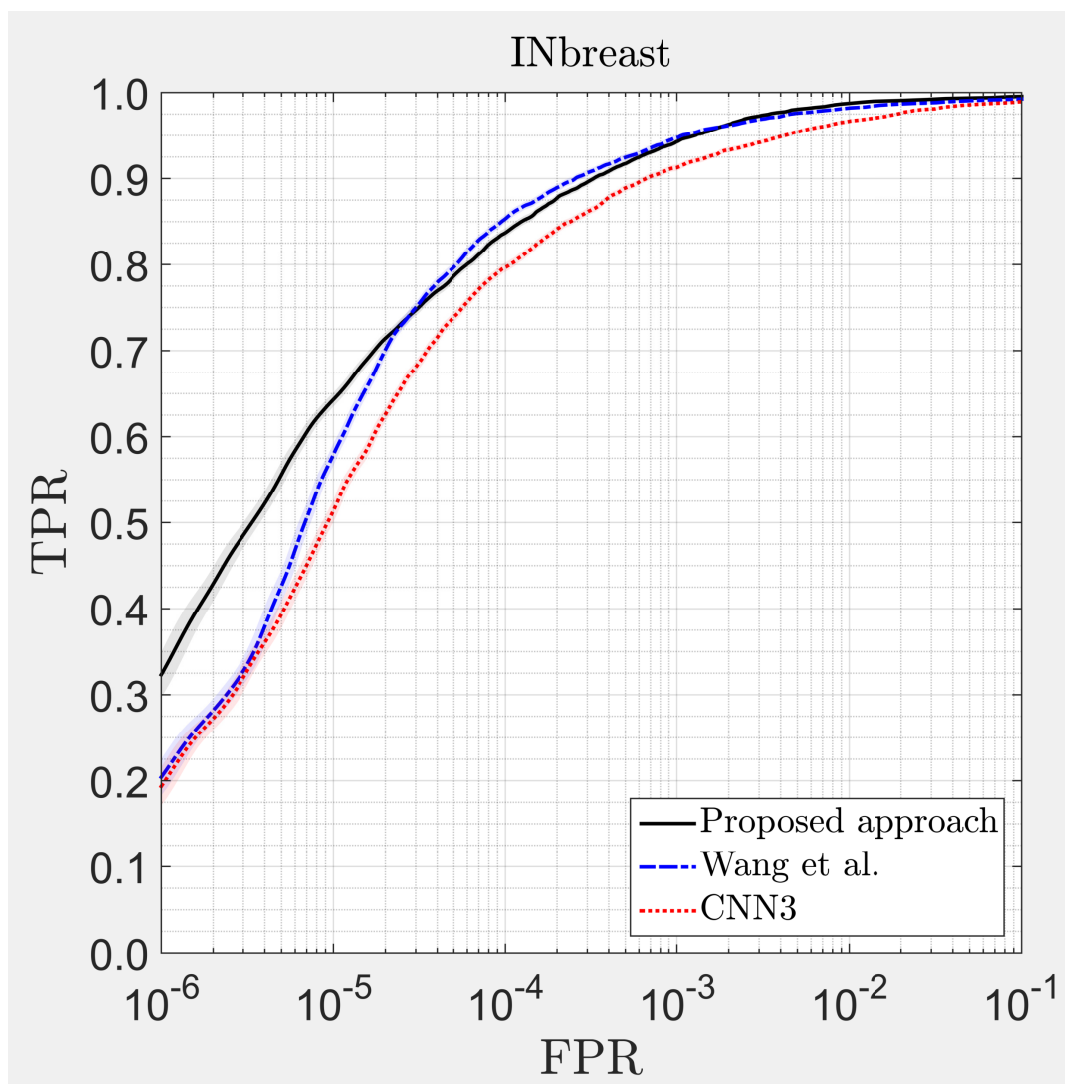


Figure 4.6: Average ROC curves obtained from 1,000 bootstrap iterations for INbreast dataset. Confidence bands indicate 95% confidence intervals along the TPR axis.

proposed by [97], a context-sensitive deep learning approach for MCs detection. To this end, the network architecture and the training settings reported in [97] were faithfully reproduced and its performance were evaluated in terms of mean sensitivity. Specifically, the detector network proposed by [97] is formed by two subnetworks, one for extracting the local image features and one for learning the background. The two subnetworks are jointly trained and the image features from the two branches are fed together into the fully-connected layers for classifying whether the input object is an MC or not. For the sake of completeness, we also compared these two approaches with the best single CNN, that is CNN3. Statistical comparisons were performed by means of bootstrapping [75]. On the test set, average ROC curves were calculated over 1,000 bootstraps and are reported in Fig. 4.6. The ROC curves of the proposed context-sensitive ensemble were notably higher in the FPR range of major interest with respect to those obtained from the other approaches.

Table 4.6: Comparative results of mean MC detection sensitivity  $\bar{S}$ 

Method	$\bar{S}$	Compared to	Difference	$p$ -value
CNN3	77.45	-	-	-
[97]	80.84	-	-	-
Proposed approach	<b>83.54</b>	CNN3	+6.09	< <b>0.025</b>
		[97]	+2.7	< <b>0.025</b>

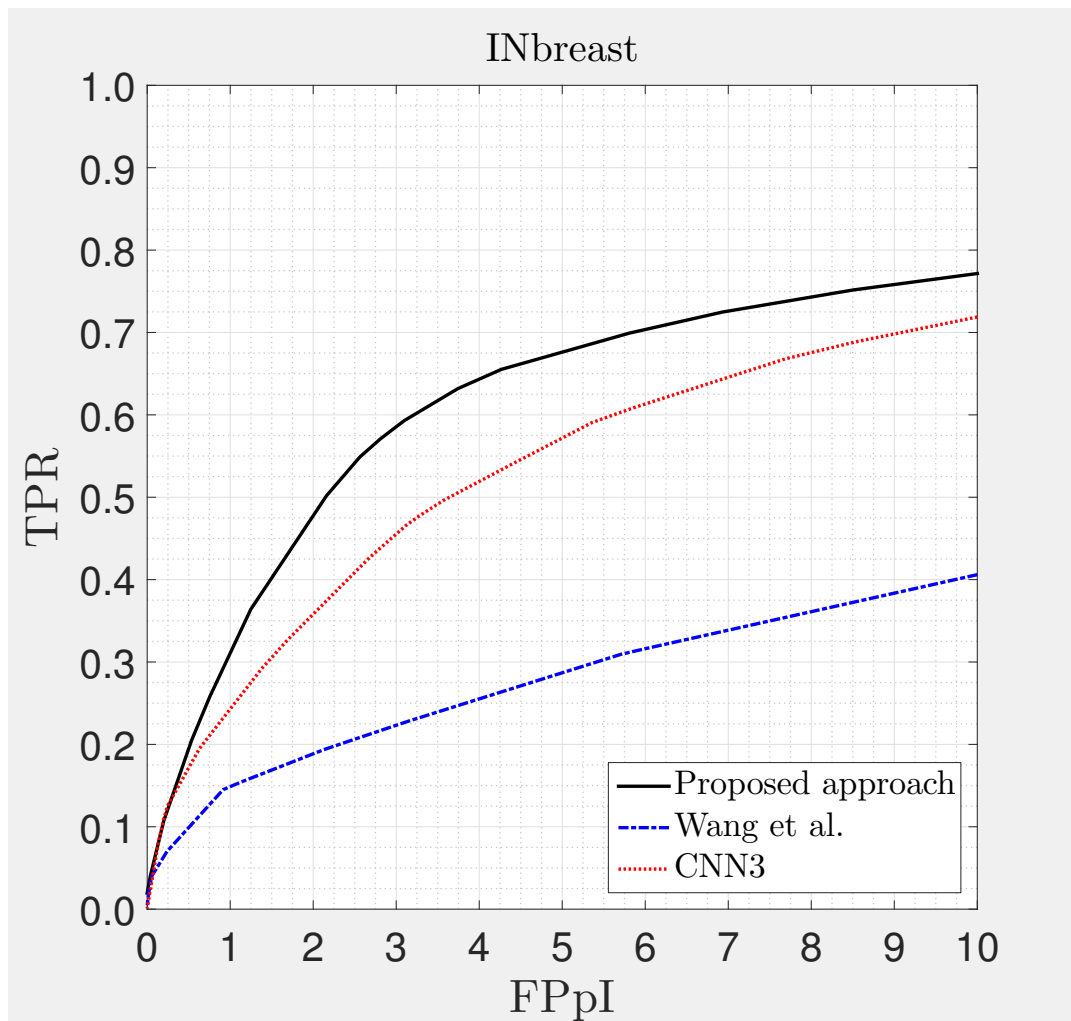


Figure 4.7: FROC curves for (a) INbreast dataset

Additionally, the mean sensitivity was calculated for each bootstrap and  $p$ -values were computed for testing significance. The statistical significance level was chosen as  $\alpha = 0.05$ , but performance differences were considered statistically significant if  $p < 0.025$  due to the Bonferroni correction<sup>1</sup> [76]. Comparative results are reported in Table 4.6. Results of the proposed

<sup>1</sup>the significance level was obtained as  $\alpha$  divided by the number of comparisons

## CHAPTER 4. Multi-context ensemble of CNNs for improving the automated detection of individual microcalcifications

Table 4.7: Comparative results of the FROC score and sensitivities at specific FPpI

Dataset	Method	Sensitivity against FPpI							FROC score
		1/8	1/4	1/2	1	2	4	8	
INbreast	CNN3	0.0699	0.1260	0.1684	0.2378	0.3531	0.5148	0.6746	0.3064
	[97]	0.0515	0.0713	0.0990	0.1487	0.1886	0.2552	0.3611	0.1679
	Proposed approach	0.0684	0.1186	0.1925	0.3025	0.4695	0.6433	0.7430	<b>0.3625</b>

Table 4.8: Results of MC per-image processing time for the trained networks

Method	$t_{MC}$
CNN1	7 s
CNN2	21 s
CNN3	78 s
CNN4	280 s
Wang et al. [97]	822 s
Proposed approach	386 s

architecture were statistically significantly better than the other considered approaches. The improvements in mean sensitivity were large +2.70% with respect to the context-sensitive approach of [97], and with respect to the best standalone CNN, +6.09% , revealing to be significantly better in detecting lesions.

To assess the performance on the whole image, the lesion-based FROC curve was calculated. Being  $r$  the radius of a lesion in the ground truth, a detected region is considered as a TP if its distance to the centre of a true lesion is no larger than  $r$ ; otherwise it is counted as an FP. To easily compare the different methods, the detection performance was summarized in a single score (FROC score) obtained by averaging the sensitivity values corresponding to the FPpI rates values of 1/8, 1/4, 1/2, 1, 2, 4, and 8, as described in [102]. Lesion-based FROC curves evaluated on the test set are shown in Fig. 4.7 for MCs detection and the relative FROC scores are reported in Table 4.7. The performance of the proposed ensemble is notably higher than the others, proving the effectiveness of the proposed method also when applied on the whole image.

Finally, per-image processing times are reported in Table 4.8. As expected, the time needed for testing a single image increases with the input size, being strictly related to the network depth. The testing time of the proposed approach is evaluated as the sum of the processing time of the 4 standalone CNNs, resulting to be lower than the time required by [97].

## 4.4 Discussion and Conclusions

In this thesis, a novel and effective method for the detection of individual MCs in digital mammograms is proposed, as a result of an analysis of the limitations of the current methods proposed for similar applications.

First, I investigated the performance of CNNs when using larger image windows during the training phase together with deeper architecture. The obtained results indicate that using small patches, hence focusing only on the local image characteristic of a lesion, is not sufficient to obtain high detection performance. This is because, ignoring the context in which the lesions are in, the detector response is susceptible to all the lesion-like image patterns that lies in the background, affecting the overall performance of the classifier. However, even too large image patches are not sufficient to obtain high level performance, being the network not able to capture the fine details of the lesions and to recognize them in their broad spectrum of appearance. Moreover, deeper networks are more difficult to train, due to the vanishing signals and the internal covariate phenomenon.

To include both local and larger contextual information, we decided to combine at inference time the predictions coming from the individual networks, resulting in the proposed multi-context CNN ensemble. The ensemble combines the predictions of 4 different networks, each one with a different level of depth and processing at training time input patches with a different level of context-information. The devised approach achieved statistically significantly more accurate results in detecting small lesions when compared to standalone CNNs, and it additionally outperformed the context-sensitive approach proposed by [97] for similar tasks.

The obtained results proved the effectiveness of using different pathways, where each path specializes in capturing information at different context levels so that the system is able to closely learn the global contextual features as well as the local detailed features. The local appearance of the lesions and their underlying characteristics were captured, with different specificity levels, by the first two pathways, while higher level features, such as the nature of the tissues of the lesions are inserted in, were learned by the deeper paths. As a result, we obtained a set of specialized and complementary detectors (based on different representations derived from the different contexts) whose combination led to a final system that is able to overcome the limitations of single-pathway networks, with a clear improvement of the discriminating power. Moreover, the reported results suggest that the approach of training the networks separately and averaging the outputs at inference time is effective to get over the optimization difficulties that might occur in the case of joint training. We think that, when the multiple pathways are simultaneously trained as in [97], the detector might find it difficult during the learning phase to come across the co-adaption between the local and the global pathways. We believe that the improvements in detection performance obtained in this work can be transferred to full CAD schemes, including diagnosis modules which can determine the nature of the detected lesions. These systems could be implemented in a routine clinical setting, being very useful to the clinicians not only for detecting suspect cases, but also for assisting in the diagnostic decision as a second reading.

**CHAPTER 4. Multi-context ensemble of CNNs for improving  
the automated detection of individual microcalcifications**

---



## Chapter 5

# Computer aided detection and diagnosis of clustered microcalcifications

---

Traditionally, detection and diagnosis of clustered MCs are separately performed and in turn the detection process is performed into two steps. In the first step, an MC detector is applied to locate the candidates of individual MCs in a mammogram: it is called candidate detector and its primary goal is to greatly reduce the number of search locations while achieving a sensitivity near 100%. In the second step, detected MCs are grouped into clusters according to a set of clustering criteria [103]. The starting point for classification methods are hence clustered MCs that are subsequently classified into being malignant or not [104]. As previously said, while successful in achieving high sensitivity, these methods are usually affected by the frequent occurrence of false positives [105], limiting the benefit that the CAD system can provide. Nevertheless the era of traditional CAD might be coming to an end, due to the rise of the new type of systems based on high-accuracy artificial intelligence that are rapidly closing the gap between humans and computers for many applications [13, 14, 15]. As a consequence a new generation of Computer-Aided Detection and Diagnosis systems has started with new solutions and perspective for digital mammograms tools [16, 17].

Recently, Wang et al. [105] developed a deep neural network to directly detect the presence of clustered MCs in mammograms. Instead of first detecting the MCs individually, they employed a convolutional neural network (CNN) to determine directly whether an image region contained an MC cluster or not. This detector was intended only for identifying suspicious regions for later examination. Subsequently the same authors [97] extended their work, incorporating local MC features so that to improve the accuracy in detecting individual MCs: they proposed a context-sensitive DNN by merging, at training time, features coming from two different CNNs, one operating on the local image features of MCs and the other on the surrounding image background. Hou et al. [106] proposed an unsupervised one-class deep convolutional autoencoder by using only image from normal subjects, to detect MC clusters. Cheng et al. [107] proposed a two-step CNN methods for detecting and classifying MC clusters. A first DCNN was trained to

discriminate clusters on individual MCs candidate detection; afterwards a second DCNN was used to provide cancer likelihood prediction. Lotter et al. [17] proposed a two-stage curriculum learning-based approach for mammograms classification, by exploiting informations coming from detected cancer lesions (MCs). They first trained patch-level CNN classifiers at multiple scales for lesion detection, which were then used as feature extractors in a sliding-window fashion to build an image-level model and render a decision on the whole image. Although deep learning methods can be successfully used for detection and classification separately, it has been shown that the best gains are obtained when systems are trained end-to-end [108]. This is achieved by training the network with a multi-task loss that uses shared representation among the related tasks to enable the model to generalize better on each original task [109, 110, 111]. Based on that, we decided to investigate the feasibility of detecting and classifying MCs cluster simultaneously by using a single end-to-end trainable network. Recently, few other works tried to combine multiple tasks in medical image analysis field: in particular for digital mammograms, Al-Masni et al. [112] proposed to use a ROI-based CNN, the so-called YOLO network, to detect and classify breast masses.

In this thesis an end-to-end system is presented to combine both the candidate detector and the classification step into one single model, and additionally segment malignant lesions in digital mammograms. The proposed architecture is a modified U-Net in which the typical encoder-decoder pathway is used to segment single MCs and address the detection problem, while a second additional branch in the bottleneck enables the network to perform classification.

The rest of the chapter is organized as follows: in Section 5.1 multi-task learning is introduced together with a detailed description of the proposed approach in Section 5.2. In Section 5.3 experimental analysis is reported along with experimental results in Section 5.5. In Section 5.6 conclusions are drawn and future research directions are outlined.

### 5.1 Multi-task learning

In Multi-task learning, there are multiple learning tasks each of which can be a general learning task such as supervised tasks (e.g., classification or regression problems), unsupervised tasks (e.g., clustering problems), semi-supervised tasks, reinforcement learning tasks, multiview learning tasks or graphical models. Among these learning tasks, all of them or at least a subset of them are assumed to be related to each other. In this case, it is found that learning these tasks jointly can lead to performance improvement. This observation leads to the birth of Multi-Task Learning (MTL) that can be defined as a learning paradigm whose aim is to leverage useful information contained in multiple related tasks to help improving the generalization performance of all the tasks. MTL is inspired by human learning activities where people often apply the knowledge learned from previous tasks to help learn a new task. Similar to human learning, it is useful for multiple learning tasks to be learned jointly since the knowledge contained in a task can be leveraged by other tasks. In a machine learning perspective, multi-task learning can be seen as a form of inductive transfer. Inductive transfer can help improve a model by introducing an inductive bias, which

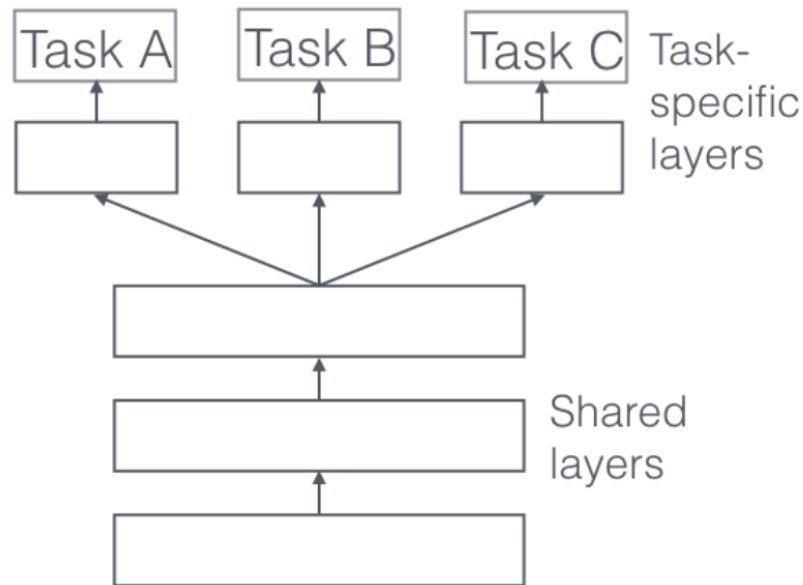


Figure 5.1: Hard parameter sharing for multi-task learning in deep neural networks

causes a model to prefer some hypotheses over others. For instance, a common form of inductive bias is  $l_1$  regularization, which leads to a preference for sparse solutions. In the case of MTL, the inductive bias is provided by the auxiliary tasks, which cause the model to prefer hypotheses that explain more than one task. In the context of deep learning, multi-task learning is typically done with either hard or soft parameter sharing of hidden layers.

### 5.1.1 Hard parameter sharing

Hard parameter sharing is the most commonly used approach to MTL in neural networks and was firstly introduced in [113]. It is generally applied by sharing the hidden layers between all tasks, while keeping several task-specific output layers (see Fig. 5.1). Hard parameter sharing greatly reduces the risk of overfitting. In fact, in [114] Baxter et al. showed that the risk of overfitting the shared parameters is an order  $N$ , where  $N$  is the number of tasks, smaller than overfitting the task-specific parameters, i.e. the output layers. This makes sense intuitively. The more tasks we are learning simultaneously, the more our model has to find a representation that captures all of the tasks and the less is our chance of overfitting on our original task.

### 5.1.2 Soft parameter sharing

In soft parameter sharing on the other hand, each task has its own model with its own parameters (see Fig. 5.2). The distance between the param-

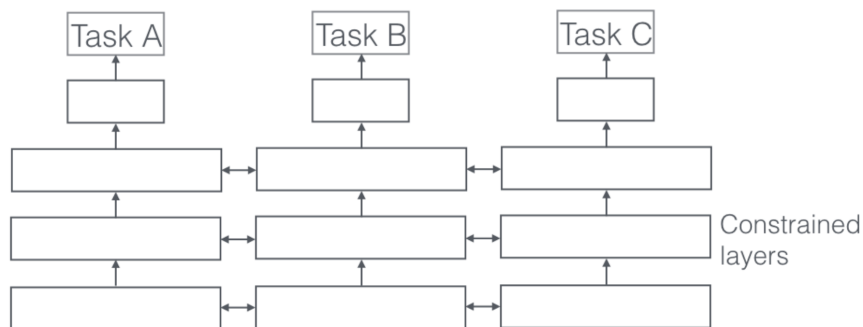


Figure 5.2: Soft parameter sharing for multi-task learning in deep neural networks

eters of the model is then regularized in order to encourage the parameters to be similar. For instance in [115] the norm is used for regularization, while in [116] the trace norm is used.

### 5.1.3 Mechanism underlying MTL

Assuming to have two related tasks, A and B which rely on a common hidden layer representation F, the mechanisms that underlie MTL can be summarized as follows:

#### Implicit data augmentation

MTL effectively increases the sample size that we are using for training our model. As all tasks are at least somewhat noisy, when training a model on some task A, our aim is to learn a good representation for task A that ideally ignores the data-dependent noise and generalizes well. As different tasks have different noise patterns, a model that learns two tasks simultaneously is able to learn a more general representation. Learning just task A bears the risk of overfitting to task A, while jointly learning A and B enables the model to obtain a better representation through averaging the noise patterns.

#### Attention focusing

If a task is very noisy or data is limited and high-dimensional, it can be difficult for a model to differentiate between relevant and irrelevant features. MTL can help the model to focus its attention on those features that actually matter as other tasks will provide additional evidence for the relevance or irrelevance of those features.

#### Representation bias

MTL biases the model to prefer representations that other tasks also prefer. This will also help the model to generalize to new tasks in the future, as a hypothesis space that performs well for a sufficiently large number of

Table 5.1: Details of the *classification block*

Layer	Type	Output size	Kernel Size	Stride
1	BatchNorm	$1024 \times m \times m$		
2	ReLU	$1024 \times m \times m$		
3	Convolutional	$1024 \times m \times m$	$3 \times 3$	2
4	BatchNorm	$1024 \times m \times m$		
5	ReLU	$1024 \times m \times m$		
6	Convolutional	$1024 \times m \times m$	$3 \times 3$	2
7	Convolutional	$1024 \times 1 \times 1$	$1 \times 1$	1

training tasks will also perform well for learning novel tasks as long as they are from the same environment.

### Regularization

Finally, MTL acts as a regularizer by introducing an inductive bias. As such, it reduces the risk of overfitting as well as the complexity of the model, i.e. its ability to fit random noise.

## 5.2 Proposed approach

In this thesis, a novel MCs cluster detection and classification method is presented as based on a single fully convolutional neural network (FCN) that performs detection and classification concurrently by using MTL. Cluster detection is treated as a segmentation problem where individual calcifications are detected, by also providing information about the location and shape of single regions. A classification branch is used to predict the presence of malignant clusters, by exploiting at the same time the information coming from the encoding path and the segmentation map.

The rationale is that, segmenting individual MCs can provide fine details about the cluster configuration, resulting in a more accurate cluster detection process. Moreover in a CADx perspective, accurately detecting the individual MCs in the cluster is important for classifying the cluster as being malignant or not. Many studies have shown that the accuracy of detected individual MCs can impact on the CADx performance [117, 118, 119]. As a consequence, the classification branch will benefit from the segmentation process in giving prediction about the malignancy of the clusters that are potentially detected in the segmentation step. On the other hand, the classification branch will positively affect the detection of MCs clusters, since the information about the category of the image will help to reduce the searching space of the MCs detection model, and hence will facilitate the cluster localization process.

### 5.2.1 Network architecture details

As mentioned above, the idea behind the proposed method is to treat detection and classification of MCs cluster as related problems. To accomplish

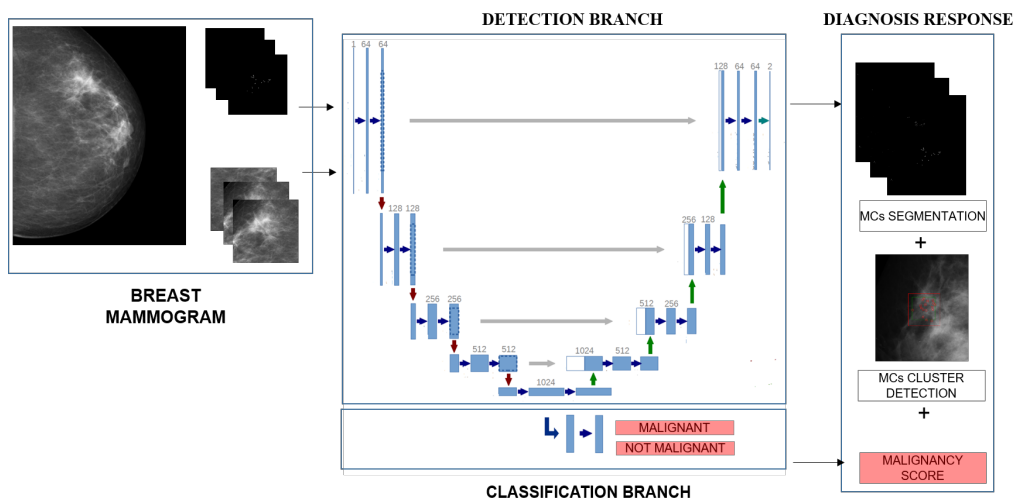


Figure 5.3: Overview of the proposed method

these two tasks simultaneously the proposed network consists of two different paths each one for each of the two defined tasks (see Fig. 5.3).

The first path is defined for the scope of MCs cluster detection and is performed by segmenting individual calcifications. For that reason we decided to follow the architecture of a basic U-Net, the most popular and successful deep learning model in the domain of medical image segmentation [59]. The main reason behind the choice of the U-Net is in the presence of the shortcut connections 2.9 that allow to combine deep, semantic, feature maps from the decoder sub-network with shallow, low-level feature maps from the encoder one, resulting to be very effective in recovering fine-grained details of the objects and to generate segmentation masks with fine details even on complex background, like breast tissue background. In addition, to reduce the issue of internal covariate shift, speed up learning and improve performances Batch Normalization layers were added before each block of the basic U-net for feature map normalization [120]. Batch Normalization layers are in fact used to independently normalize the feature values of each layer to zero mean and unit standard deviation during each training batch. Given that we want to perform the task of image classification together with detection an auxiliary branch was added to the original encoder-decoder path and placed at the end of the encoder so that to exploit the set of extracted features for classification. Following the structure of the encoding path, this classification branch consists of a double convolution block, ending with a final  $1 \times 1$  output convolutional layer on top of which a sigmoid function is applied for the purpose of final binary classification (see Table 5.1).

## 5.2.2 Multi-task loss

An important step in multi-task learning to address the multiple tasks simultaneously and make the model end-to-end trainable is the use of multiple losses. We design our multi-task loss as the sum of two different terms

$$L_{\text{tot}} = L_{\text{seg}} + \lambda L_{\text{class}} \quad (5.1)$$

where  $\lambda$  is a weighting factor that is meant for balancing the two losses

in terms of magnitude and enable the learning of both tasks with equal importance, without allowing one to dominate.

### Segmentation loss

For the segmentation task we define  $L_{seg}$  as a pixel-wise *weighted top-k* Cross-Entropy loss. The Cross Entropy loss is defined by the following equation:

$$CE_{loss}(\mathbf{t}, \mathbf{s}) = - \sum_{i \in \{P, N\}} t_i \log s_i \quad (5.2)$$

where  $\mathbf{s}$  is the vector of pixel-wise class probabilities produced by the networks and  $\mathbf{t}$  the vector of relative groundtruth.

However, MCs segmentation is a severely unbalanced segmentation problem since the positive class, represents a very small percentage of the total pixels distribution that is dominated by negative pixels belonging to the image background. A common recent method to address class imbalance and avoid the learning being overwhelmed by the majority class is to introduce different weighting factor  $\alpha_i$  for the two classes. For this reason we use a *weighted* Cross Entropy loss by giving higher weights to positive pixels.

$$WCE_{loss}(\mathbf{t}, \mathbf{s}) = - \sum_{i \in \{P, N\}} \alpha_i t_i \log s_i \quad (5.3)$$

However if the weights are used to balance the importance of positive and negative pixels, they do not make any differentiation between easy/hard examples. Since our segmentation problem is heavily skewed it is reasonable to think that the majority of easy pixels belongs to the background, as opposed to a small number of hard positive examples. Thus we propose to use as final loss for segmentation a *top-k pixel-wise weighted* Cross Entropy loss

$$L_{seg} = L_{top-k}(WCE_{loss}) \quad (5.4)$$

The proposed loss consists in selecting the top  $k$  percentage of the most difficult pixels, that is the pixels with the highest cross-entropy loss and only add their contribution to the total loss. In this way it is possible to focus training on positive pixels, discarding the majority of easily classified negatives.

### Classification loss

For the classification task *Focal loss* is applied, being a way to focus training on hard patches, by suppressing easy ones.

$$F_{loss}(t, s) = - \sum_{i \in \{P, N\}} (1 - s_i)^\gamma t_i \log s_i \quad (5.5)$$

As shown in equation [5.5](#) *Focal loss* adds a modulating factor  $(1 - s_i)^\gamma$  to the Binary Cross Entropy loss, that is meant to leave the loss value unaffected when samples are misclassified and to down-weight it when they are correctly classified. In addition the parameter  $\gamma$  is designed to tune the modulating factor and adjusts the down-weight rate.



Table 5.2: Distribution of the digital mammography (DM) exams included in this study.

	DM studies	DM images	samples
normal	584	1745	27,929
malignant	236	450	7,200

### 5.2.3 Online Hard Example Mining

The final goal of the proposed method is to detect and classify MCs clusters that is again a heavily unbalanced problem since the number of normal images is higher than those containing MCs. Thus, beyond addressing the class skew between positive and negative pixels with the multi-task loss, we also need to feed the network with a sufficient number of malignant clusters. To this end, we applied Online Hard Mining (OHAM) by performing hard example selection epoch-wise. Given a set of samples of size  $M$  we performed regular forward propagation and assign each sample a weight  $w_i$  whose magnitude is inversely related to the ability of the network to segment and classify that particular sample, according to the values of the performance measures. The hard mining rule can be defined by the following expression:

$$\mathbf{W}_M = (\beta_P \mathbf{W}_{M_P}, \beta_N \mathbf{W}_{M_N}) \quad (5.6)$$

where  $\mathbf{W}_{M_P}$  and  $\mathbf{W}_{M_N}$  are the weight vectors for all the positive and negative samples respectively. The  $k$ -th element of these vectors are given by:

$$W_{M_i,k} = \frac{w_k}{\sum_{k=1}^{M_i} w_k} \quad \forall k = 1, \dots, M_i \quad \text{and} \quad i \in \{P, N\} \quad (5.7)$$

At the end of  $n$  epochs a resampling is done so that samples with higher weights are fed more often to the network during the subsequent epochs. The OHAM is separately done for positive and negative samples with the weights  $\beta_P$  and  $\beta_N$  that sum up to 1 and are designed to guarantee that a certain percentage of samples belonging to positive/negative classes is selected for the next epochs.

## 5.3 Materials

### 5.3.1 Dataset

This study was conducted with anonymized data, retrospectively collected from DIAG institutional archive. The study was approved by the regional ethics board after summary review, with waiver of a full review and informed consent. All cases with biopsy-proven malignant lesions were collected, while normal exams were selected if they had at least two years of negative follow-up. This yielded a total of 820 exams, from which 236 exams contain biopsy-verified malignant MCs clusters. Most exams were bilateral and included two views (CC and MLO), resulting in a total of 2195 images. The exact distribution of the images is summarized in Table 5.2. The images were acquired by using two digital mammography (DM) machines from



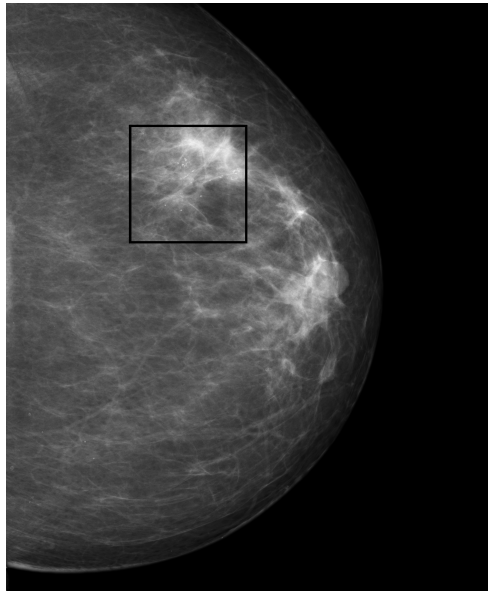


Figure 5.4: DIAG dataset annotation example

General Electric (Senographe 2000D and Senographe DS) with an image resolution of  $94\mu m$ . MCs cluster together with single MCs in each image were manually identified by a group of experts, by providing coordinates of cluster bounding boxes and the position of the individual MCs (see Fig. 5.4), that we both used as ground truth in this study.

The collection of positive mammogram images was partitioned into three subsets, one with 155 cases (300 images) for training, one with 30 cases (50 images) for validation, and one with 51 cases (100 images) for testing. To avoid any potential bias, the different views from one case were assigned together to either the training, validation, or testing subset exclusively.

### 5.3.2 Groundtruth image generation

Since the detection problem was addressed by segmenting individual calcifications, we also needed binary masks as reference standard for the segmentation step. Therefore a first segmentation process was applied on mammograms in order to obtain segmentation masks to use as references. The segmentation process was performed by exploiting informations coming from the annotation files and by using image analysis techniques like thresholding techniques and mathematical morphology.

#### Image binarization and Otsu thresholding algorithm

Binarization plays an important role in digital image processing, mainly in computer vision applications. Thresholding is an efficient technique in binarization. The choice of thresholding technique is crucial in binarization. Several thresholding algorithms have been investigated and proposed to define the optimal threshold value. The thresholding algorithms can be categorized into six classes: histogram shape-based methods, clustering-based methods, entropy-based methods, object attribute-based methods, the spatial methods and local methods based on the local characteristics of each

## CHAPTER 5. Computer aided detection and diagnosis of clustered microcalcifications

---

pixel. Otsu's thresholding technique is a classification-based method which searches for the threshold that minimizes the intra-class variance, defined as a weighted sum of variances of the two classes. Given this definition, the Otsu's method segments the image into two light and dark regions T0 and T1, where region T0 is a set of intensity level from 0 to t or in set notation  $T0 = \{0, 1, \dots, t\}$  and region  $T1 = \{t, t + 1, \dots, l - 1, l\}$  where t is the threshold value and l is the image maximum gray level. T0 and T1 can be assigned to object and background or vice versa. Otsu's thresholding method scans all the possible thresholding values and calculates the minimum value for the pixel levels each side of the threshold. The goal is to find the threshold value with the minimum entropy for the sum of foreground and background. Otsu's method determines the threshold value based on the statistical information of the image where for a chosen threshold value t the variance of clusters T0 and T1 can be computed. The optimal threshold value is calculated by minimizing the sum of the weighted group variances, where the weights are the probability of the respective groups.

### Mathematical morphology

Mathematical morphology is a framework for image processing based on lattice theory and random geometry and it is a tool for investigating geometric structures in binary and greyscale images. The theory of mathematical morphology is built on two basic image processing operators: the dilation and the erosion. Dilation allows objects to expand, thus potentially filling in small holes and connecting disjoint objects. Erosion shrinks objects by etching away (eroding) their boundaries. A feature of these operators is the fact that they preserve the objects' shapes for the most part. Morphology is thus a theory where size and shape of objects play an important role. An important parameter of the morphological operators is the structuring element. This is a small set, mostly much smaller than the image set, that scans the image. The pixels covered by this structuring element determine the output value of the currently covered pixel. The dilation of a set A by structuring element B is defined by:

$$A \oplus B = \bigcup_{b \in B} A_b \quad (5.8)$$

The erosion of a set A by structuring element B is defined by:

$$A \ominus B = \{z \in E^2 : (B)_z \subseteq A\} \quad (5.9)$$

Using these basic operators, much more complex operators can be constructed i.e opening, closing, filling that can be used in different applications.

### Groundtruth generation algorithm

For the process of groundtruth generation, starting from single point annotations,  $p \times p$  patches were generated around each single point so that to focus on region of interests and provide more accurate segmentation. The segmentation process was therefore applied on single patches and can be summarized as follows: (i) the individual calcifications were re-centered, since not all the annotated points exactly matched centers of the relative

Table 5.3: Hyperparameter tuning and optimization

Hyperparameters	Tuning range	Step size	Optimal value
Learning decay	step-cosine		cosine
Init learning rate	$10^{-5} - 10^{-2}$	10	$10^{-2}$
$\eta_{min}$ for cosine decay	$10^{-7} - 10^{-5}$	10	$10^{-5}$
$\alpha_P$ for WCE	0.1 – 0.9	0.1	0.9
$Top-k$ for $L_{seg}$	1% – 50%	10-5-1	1%
$\beta_P$ for $W_{MP}$	0.1 – 0.9	0.1	0.5
$\lambda$ for MTL	$10^{-5} - 10^{-2}$	10	$10^{-2}$

MCs; (ii) a Otsu thresholding algorithm [121] was applied in order to obtain first rough binary masks; (iii) a first refinement step was performed by means of morphological closing and filling to respectively close contours and fill the holes; (iv) a refinement of segmented region borders was done by evaluating the Euclidean distance between each border pixel and the nearest zero pixel of the binary mask and suppressing all border pixels with distance equal to zero; (v) segmentation was validated according to the values of segmentation area and eccentricity of the segmented MCs; and (vi) in case of bad segmentation a Gaussian distribution with the same axes length and orientation of the MCs resulting from the first segmentation step was generated and used as starting point to repeat the steps (ii)-(v)

### 5.3.3 Data samples extraction

In the experiments, image patches were extracted from all the mammograms to train the proposed network. The size of the image patches was chosen so as to entirely contain MCs clusters in output images. Given that the estimated 95% percentile of MCs cluster size in our image is  $\approx 4.2cm$  with an image resolution of  $94\mu m$ , an output image patch size of  $470 \times 470$  pixels was used. Since the image size is decreased during the flow through the network, an input patch size of 652 pixels was chosen. For each mammogram in the training set, positive samples were extracted from the annotated MCs clusters in the image. Starting from each marked MC cluster location, 16 image windows were randomly cropped within the annotation bounding box in order to form input samples to the network, by yielding 7,200 positive patches. Negative samples instead were extracted by randomly cropping the same number of image patches within the breast tissue of normal images and totalizing a number of 27,929 negative patches (see Table 5.2)

## 5.4 Experiments

### 5.4.1 Performance evaluation

The ability of the model in detecting MCs cluster was evaluated by performing FROC analysis on a image and a case basis. The FROC curves were generated starting from the segmentation outputs. Given a mammogram, the network produces a probability map in which each pixel represents the

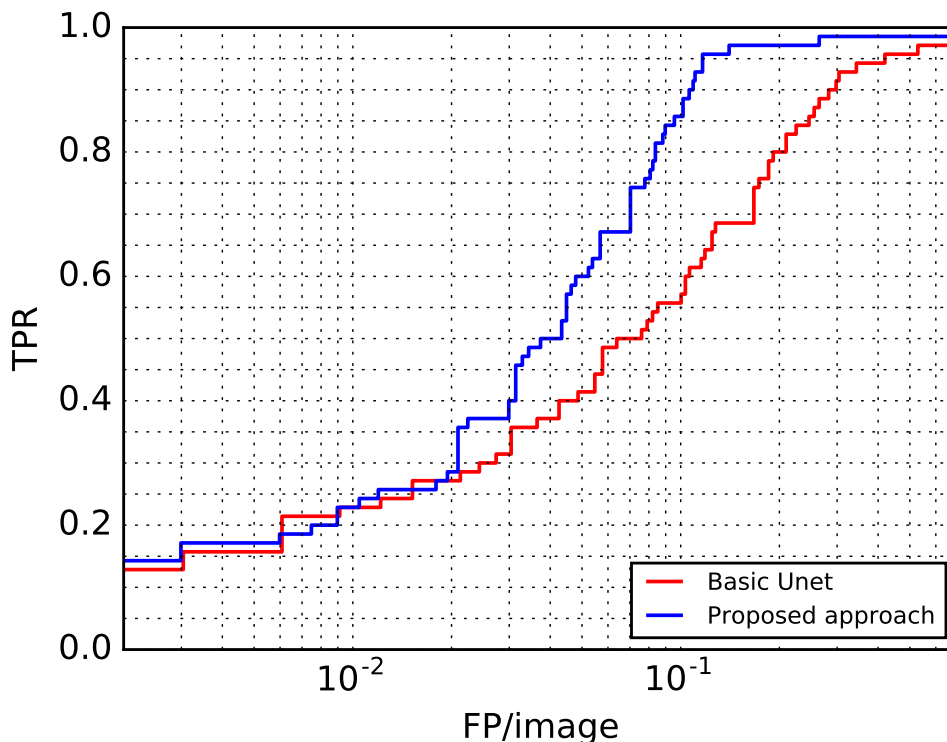


Figure 5.5: Comparison image-based FROC curves of a basic U-Net for detection and of the proposed modified U-Net for detection and classification

probability that pixel belongs to a suspicious lesion. Starting from these prediction maps, we applied binarization by thresholding at 0.5 value (which is the value which gave good performance on the validation and training sets), obtaining final segmentation maps. Once segmentation maps were generated we computed all connected components in order to have a list of all individual detected MCs. The probability values associated to individual MCs were then used to calculate a probability score, we called  $p$ -score, for each candidate detected clusters, reflecting the likelihood of the detected “object” to be a cluster or not. In the detector output, a detected object was treated as a TP when at least 40% of its area overlapped with that of a true cluster and its p-value value was greater than the threshold (where thresholds were chosen as all the p-scores associated to each candidate cluster).

The performance of the model in classifying MCs cluster into malignant or not was evaluated by means of image-based ROC curve, for a series of thresholds on the classifier output associated to each sample. The Area Under the ROC curve is also reported, being a performance measure that is insensitive to the class skew. Joint predictions between detector and classification branches were also evaluated, being a measure of the ability of the system to find and locate malignant clusters. Joint FROCs were obtained by combining detector and classification outputs.

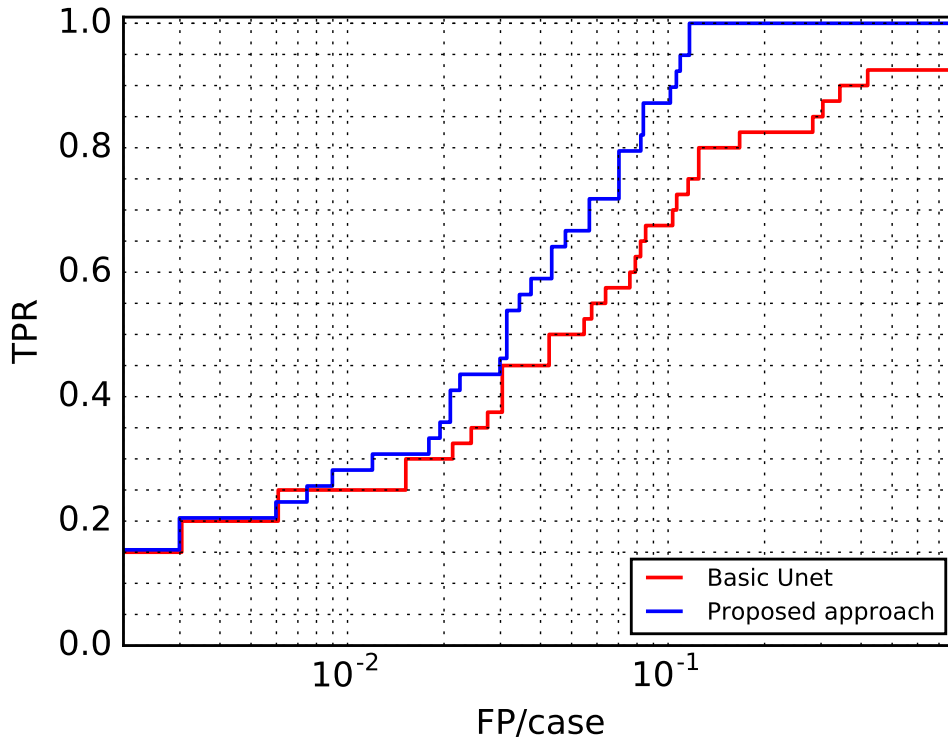


Figure 5.6: Comparison case-based FROC curves of a basic U-Net for detection and of the proposed modified U-Net for detection and classification

### 5.4.2 Model parameters

For model training, back-propagation and mini-batch stochastic gradient descent were used. The network was trained to optimize the joint loss given by Eq. 5.1. Augmentation of the two classes was performed by randomly flipping the patches horizontally and vertically and by generating new random shifting of the cropping bounding box. Since patch-based OHAM was applied, the random shifting was done by moving patches at most 20% of the patch size to guarantee that the same patches did not differ a lot trough the epochs and made OHAM effective. Different experiments with different combinations of parameters were done in order to find the model maximizing the results on both tasks. Dice similarity coefficient and classification AUC were used as primary evaluation metrics to estimate the efficiency of the network. DSC is a spatial overlap index and its value ranges from 0, indicating no spatial overlap between two sets of binary segmentation results, to 1, indicating complete overlap.

$$DSC = \frac{2|X \cap Y|}{|X \cup Y|} \quad (5.10)$$

After the network was trained the best model was chosen as the one that optimized the sum of dice score and classification AUC that were both estimated on the validation set. A complete overview of hyperparameter tuning and optimization is given in Table 5.3. The best model was finally obtained by training the network by means of a cosine annealing schedule

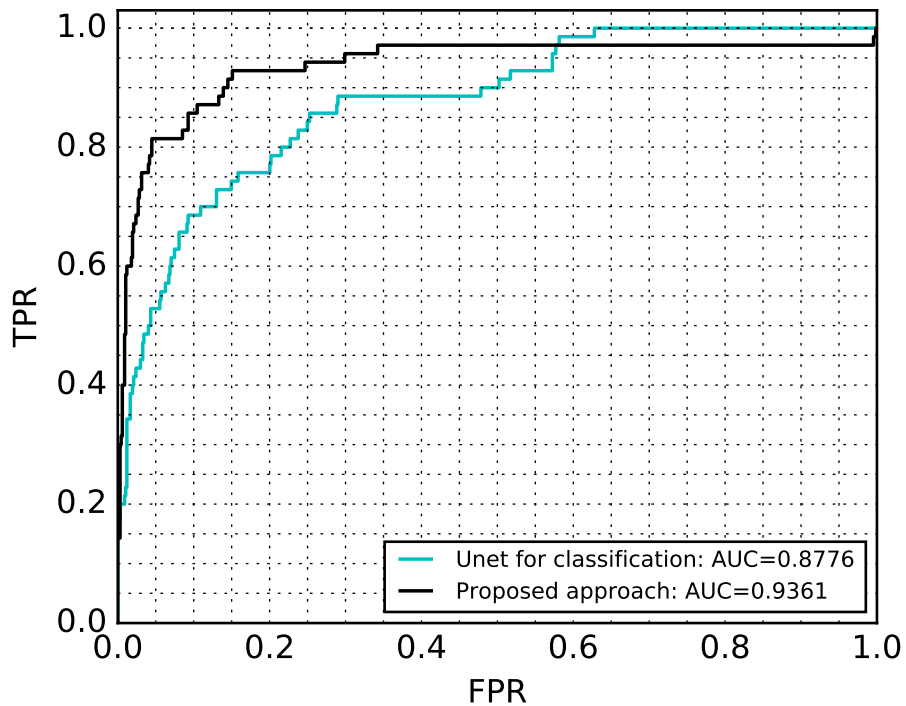


Figure 5.7: Comparison image-based ROC curves of a U-Net for classification and of the proposed modified U-Net for detection and classification

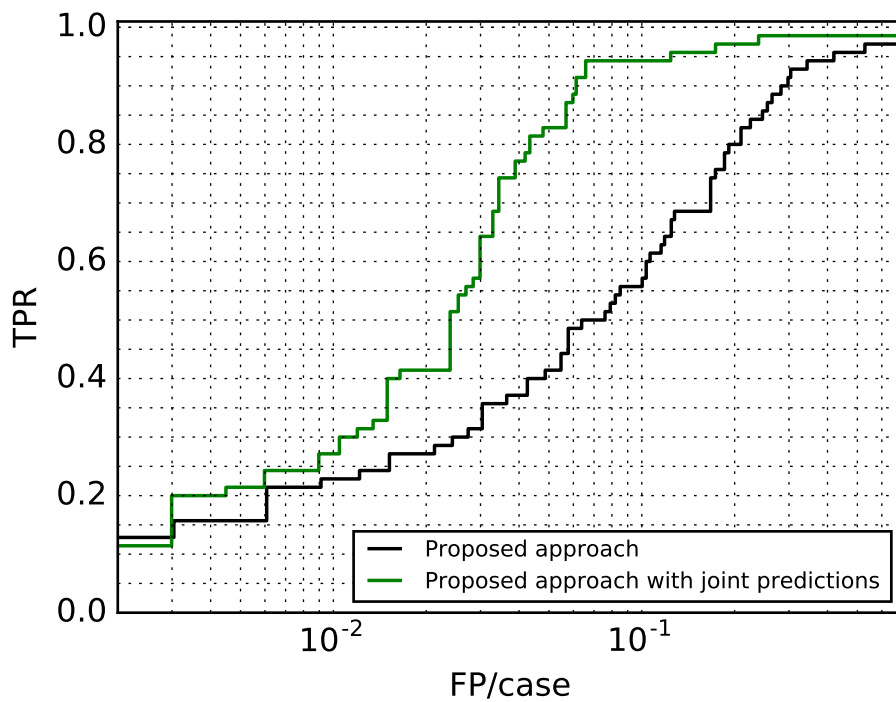


Figure 5.8: Image-based FROC curves obtained for single and joint predictions

with an initial learning rate  $l_r = 10^{-3}$  and a minimum learning rate  $\eta_{min} = 10^{-5}$ , *momentum* = 0.9 and a mini-batch size = 2.

## 5.5 Results

### 5.5.1 Cluster detection

From the experimental point of view, we first trained a stand-alone U-Net for the only task of image segmentation. The network architecture was set to a basic U-Net and hyper parameters were tuned in order to optimize individual MCs segmentation first. The major drawback in training was due to the high class imbalance between background and foreground pixels that made it difficult to stabilize learning. We investigated different values of  $K$  (see Eq. 5.4), corresponding to the percentage of pixels (the ones with the highest cross entropy loss) that were used for back-propagating the gradient. We found the optimal value to be  $K = 1$ , that we later found to be equivalent to the percentage distribution of positive pixels in the training images. This means that the learning was carried out by only 1% of the total amount of image pixels, that we suppose being mostly positive pixels, being the majority of negative samples easily classified. As for the weights  $\alpha_i$  of WCE, we found  $\alpha_{pos} = 0.9$  and  $\alpha_{neg} = 0.1$  to be the right compromise between gradient convergence in terms of training time and performance measures. In Figures 5.5 and 5.6 image-based and case-based FROC curves for the basic U-Net are reported. In the reported figures, the y-axis represents the fraction of true MCs cluster detected (i.e., sensitivity) by the detector, whereas the x-axis represents the number of FPs detected per unit mammogram or or unit case for Fig 5.5 and 5.6 respectively. Results show that this baseline model was able to achieve 60% sensitivity at 0.1 FPpI and 70% sensitivity at 0.1 FPpC.

### 5.5.2 Cluster detection and classification

After the basic U-Net was trained, the next step in the process was to add the classification branch (on top of the encoder path) and train the two task jointly. In this case, optimization was needed for finding the right value of  $\lambda$  Eq. 5.1 that properly balanced segmentation and classification tasks. The final value was chosen  $\lambda = 10^{-3}$  as lower values made the classification learning too slow, whereas higher ones caused total loss being dominate by classification contribution.

Weights  $\beta_p, \beta_n$  for OHAM in 5.6 were also tuned, by varying the percentages of positives and negatives samples that were fed into the networks through the epochs. We found that the best configuration was the one that ensured a precise balancing between positive and negative samples. Image-based and case-based FROC curves for the proposed approach are reported together with the resulting FROCs from the basic U-Net in Fig 5.5 and 5.6 respectively. As it can be seen, in both cases the FROC curve of the modified U-Net is notably higher, hence better detection performance were obtained. In particular, in the image-based FROC with FP rate at 1 FPpI, the proposed approach achieved a sensitivity of 85%, compared to 60% for the basic U-Net. Significant improvement is proved also by the case-based FROC, in which with 1 FPpC we achieved 87% sensitivity versus 68% for



the basic U-Net. These results indicate that multi-task learning is beneficial for improving the detection accuracy of MC clusters. For the classification task we evaluated the image-based ROC curve and reported the achieved AUC in Fig. 5.7.

For the sake of completeness we also trained a U-Net with the only classification branch, in order to compare the performance of a single-task network for classification with the proposed end-to-end model. In Fig. 5.7 comparison ROC curves are reported. Results show that the proposed approach obtained a notably higher AUC (0.9361) with respect to (0.8776) obtained by the classification U-Net, proving that multi-task learning was beneficial also in giving prediction about the malignancy of detected clusters. In Figs. 5.9 and 5.10 detection results from the proposed system are reported, showing its ability in detecting, locating and classifying malignant MCs clusters. For each image the system provides a bounding box of the detected cluster, together with segmentation of individual MCs and a malignancy score, indicating the confidence degree about the malignancy. In Figs. 5.9 and 5.10 the green box represents the annotation bounding box while detection bounding box is outlined in red together with single detected MCs. In these results, the operating point was set such that the FPPi rate was equal to 0.1. In 5.9 a TP detection is reported: the cluster is accurately located and single MCs precisely identified. The corresponding malignancy score is 1, showing the certainty of the network in having identified a malignant cluster. In 5.10 a FP detection is shown. As it can be seen, even if a false positive cluster has been detected the associated malignancy score is very low, indicating the classifier is almost sure there is not a malignant cluster. Starting from this observation, we decided to investigate joint predictions and see if the combination of segmentation and classification outputs would have contributed in reducing the number of FPs and hence positively affect performances. Joint FROCs were obtained by combining detector and classification outputs: the p-values associated to each sample and obtained from segmentation masks were summed up with the classifier prediction for the same sample, and then use these joint predictions as confidence degree for the decision rule. In 5.8 the FROC curve obtained from joint predictions is reported together with the one obtained from segmentation output alone. As it can be seen, the joint curve shows a significant reduction of FPs, yielding a 95% sensitivity at 0.1 FPPi. This proves that the combination of outputs from the two tasks, enforces the correct predictions and reduces the errors, minimizing the number of FP predictions and resulting in significantly improved detection performance.

## 5.6 Conclusions

In this thesis, a novel and effective method for the detection and classification of MCs cluster in digital mammograms is presented. While recent deep learning CAD systems only addressed the detection and diagnosis tasks separately, the proposed U-Net based CAD system can handle both detection and classification at the same time using whole breast image. The method is able to locate clustered MCs in the image by segmenting individual MCs, and to provide a classification score that predicts the cluster malignancy. This is accomplished by using a multi-task loss that exploits shared representation among the related tasks enabling the model to better handle



both detection and classification. The multi-task loss is made up of two different losses, specifically designed to face each single task. The problem of heavy class skew between MCs and non-MCs pixels in the segmentation task was addressed by defining a novel top-k ranking loss. As for the cluster classification, focal loss was applied for the same reason of addressing the high class imbalance between cluster and not cluster samples. Online hard mining was also used to focus learning on hard samples by discarding the majority of easily classified data. The performance of the proposed method was evaluated both on detection and classification of clustered MCs. The performance of the proposed model in detecting microcalcification clusters were compared with the one of a basic U-Net, and in the same way the effectiveness of the proposed method in classifying detected clusters was proved by comparing results with the one of a U-Net with the only classification branch. Obtained results show that in both cases multi-task learning is beneficial with significantly improved performances with respect to single-task systems. On one hand, the classification branch positively affects the detection of MCs clusters, and equivalently the classification branch takes advantages from segmentation of individual calcifications in giving predictions about the malignancy of the cluster.

Finally the joint action of detector and classifier was analysed and showed that combining predictions from the detection and the classification paths can significantly reduce the number of false positive, improving overall performance. Such a system, when applied to a clinical setting, would help the radiologists to reduce the number of unnecessarily recalled women with microcalcification clusters, thus improving the effectiveness of screening and diagnosis processes.

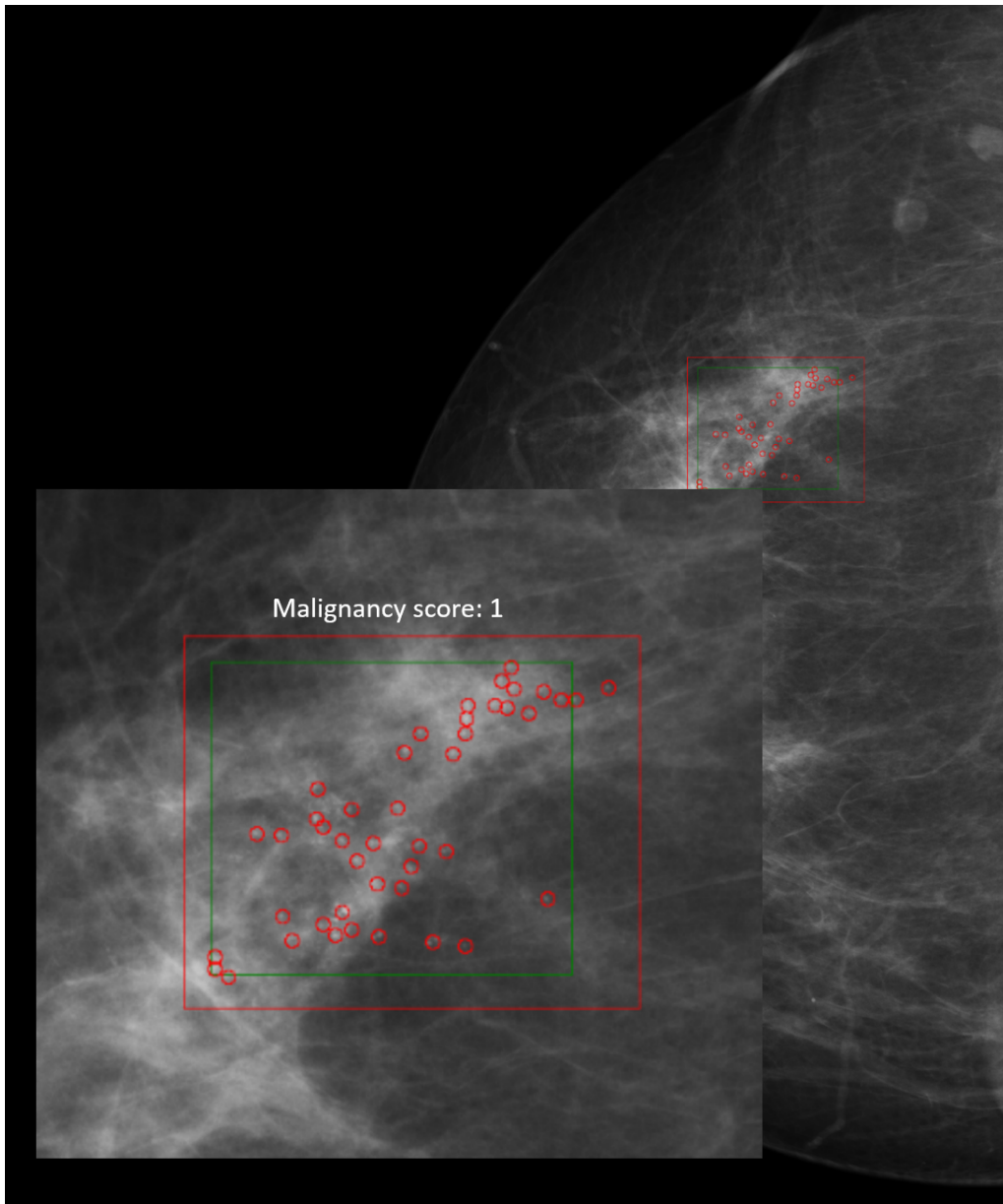


Figure 5.9: Example of a True Positive detected cluster

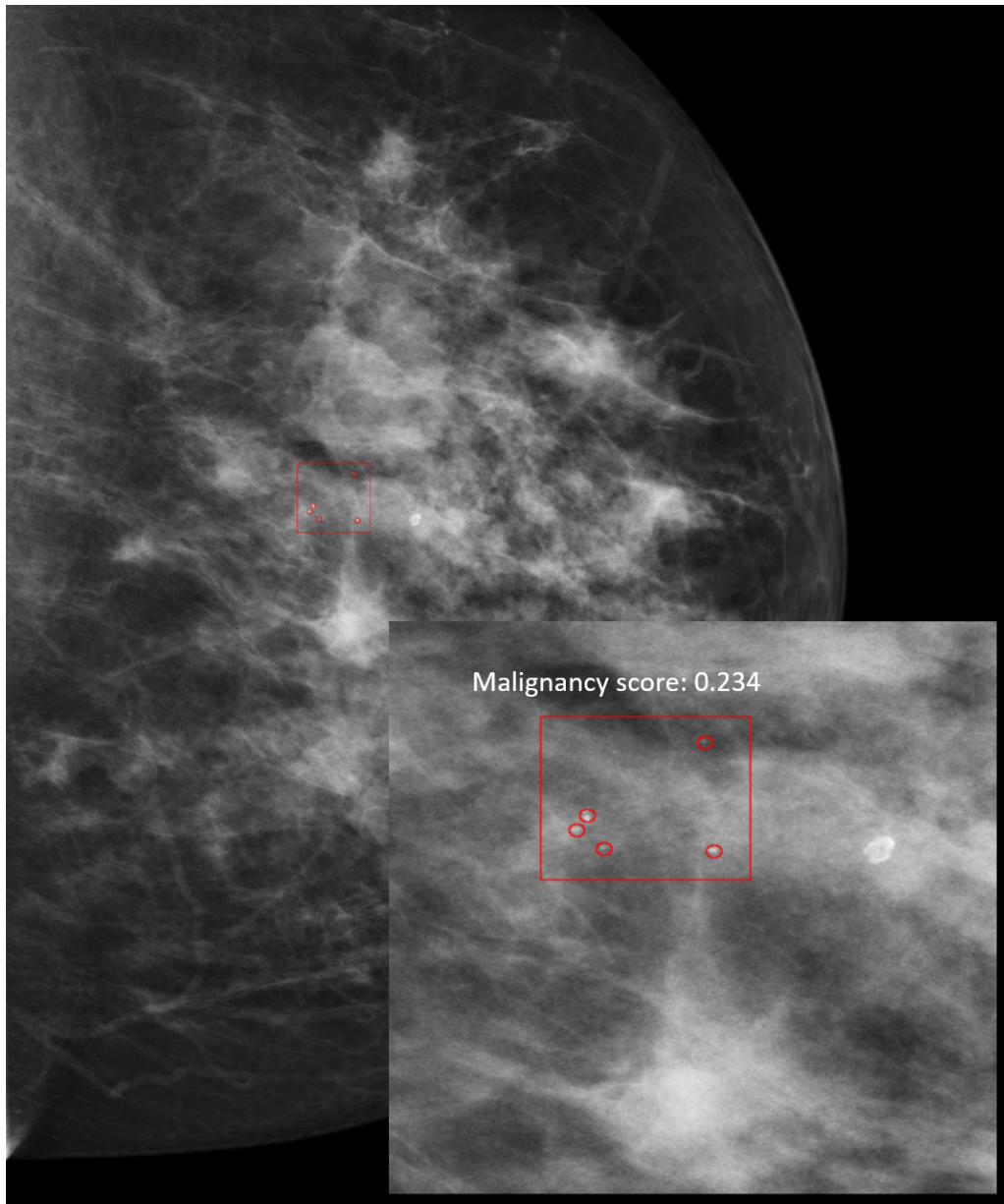


Figure 5.10: Example of a False Positive detected cluster

## CHAPTER 5. Computer aided detection and diagnosis of clustered microcalcifications

---

# Chapter 6

## Summary and Conclusions

Breast cancer is the most common cause of cancer death in women worldwide. Early detection with correct diagnosis is extremely important to increase the survival rate. In most western countries, screening programs are organized in order to detect breast cancers at an early stage. The large number of acquired screening mammograms are interpreted by radiologists, who look for mammographic indicators of cancer like clusters of microcalcifications and masses. However, interpreting screening mammograms is a big challenge even for an expert radiologist since the low prevalence makes finding abnormalities difficult and because of the low visibility and variability in the appearance of the lesions. Over the past decades several strategies have been adopted to improve breast cancer detection and among them is the development of Computer Aided Detection or Diagnosis systems, that are meant to assist radiologists in finding and locating abnormalities on the images and supporting their diagnosis response and a lot of solutions and methodologies have been proposed in the last decades. Even though traditional CAD systems show a sensitivity similar to radiologists, there are still a few hundred, false positives for every true positive in a screening setting, limiting the benefit that the CAD system can provide. However the success of novel machine learning algorithms based on deep learning convolutional neural network has enabled a new era for CAD system. The aim of this thesis was to develop a full computer aided detection and diagnosis system for clustered microcalcifications that overcame the limitations of traditional CAD system, able to reduce the gap between CAD systems and radiologists in terms of false positives while maintaining an high sensitivity standard. This has been done after an in-depth analysis of the limitations of the existing methodologies and by exploiting the advantages of novel deep learning algorithms.

In **chapter 3** the problem of detecting individual microcalcifications was addressed. The problem was treated as a classification problem in which individual pixels of the mammogram are classified as belonging to a microcalcification (i.e., to the positive class) or not (i.e., to the negative class). Like in other medical image analysis applications dealing with computerized detection of lesions, the vast majority of image locations do not contain the searched objects, and this results into an imbalance between the positive and the negative classes. As a consequence, learning an effective classifier is very difficult. To address this the thesis a two stage classification scheme was proposed, that combines the benefits of two powerful methods detecting microcalcifications, Deep Cascade and Deep Convolutional Neural Network.

## CHAPTER 6. Summary and Conclusions

---

Experiments were done on a database of 1,066 mammograms acquired with GE Senographe systems from the Radboud University Medical Center (Nijmegen, The Netherlands) after referral in screening. Results showed the effectiveness of the method, whose performance were statistically significantly higher than Deep Cascade and CNN alone, and yielded an average improvement in mean sensitivity of 3.19% and 2.45% respectively.

**Chapter 4** still faced the problem of detecting individual calcifications by focusing on the importance of image context for the task of accurately detecting small lesions. To this aim a multi-context ensemble of Convolutional Neural Networks was proposed, aiming at learning different levels of image spatial context. The main innovation behind the proposed method is the use of multiple depth CNNs, individually trained on image patches of different dimensions and then combined together. In this way the final ensemble was able to find and locate MCs on the image by exploiting both the local features and the surrounding context of the lesion. Experiments were done on the publicly available datasets, INbreast. Statistically significantly better detection performance were obtained by the proposed ensemble with respect to other approaches in the literature, demonstrating its effectiveness in the detection of MCs.

**Chapter 5** brings the gap between the detection of individual microcalcification and the detection and diagnosis of microcalcification clusters. For this purpose a novel multi-task learning system was presented, that is able to simultaneously detect and classify MCs cluster by means of a modified Unet. Individual MCs are accurately segmented by the encoder-decoder path while a classification branch in the bottleneck is meant to predict the malignancy of the potentially detected cluster. The network was trained by using a multi-task loss that uses shared representation among the related tasks enabling the model to better handle both detection and classification. The system was tested on a database of digital mammogram exams collected from DIAG institutional archive. The performance of the proposed model in detecting microcalcification clusters was compared with the one of a basic Unet, and in the same way its effectiveness in classifying detected clusters was proved by comparing results with the one of a Unet with the only classification branch. Obtained results show that in both cases multi-task learning is beneficial with significantly improved performances with respect to single-task systems. On an image level, a sensitivity of  $\simeq 0.85$  at 0.1 false positives (FP) per image was achieved together with a classification AUC = 0.9346, proving the ability of the method in effectively detect and classify MCs clusters by reducing the number of FPs, that was the main goal of this thesis. Such a system, if applied to a routine clinical environment, would be of significant help to the radiologists in significantly reduce the number of unnecessarily recalled women, thus improving the effectiveness of screening and diagnosis processes.

For future work, there are different directions that can be taken. Firstly, other network architectures can be employed for facing the multi-task learning problem e.g. Faster R-CNN [122] and Mask R-CNN [123]. Faster R-CNN is meant to face classification and bounding box regression problems where Mask R-CNN adds a branch in parallel to the existing branches in [122] to also predict segmentation masks in a pixel-to-pixel manner. Secondly, a next step could be the inclusion of the cancer classification into stages so as to provide the severity of malignancy associated to a lesion, which could further help the radiologists in the diagnostic decision. More-

---

over, the majority of the methods proposed in this work, even though designed for the microcalcification detection task, could be more generally applicable to other small lesion detection and classification problems in medical image analysis such as the automated detection of retinal microaneurysms in digital color fundus images. This would further contribute to the growth that deep learning has brought in the application of machine learning techniques for solving medical problems.





# Bibliography

- [1] P Boyle and B Levin. *World Cancer Report 2008*. International Agency for Research on Cancer, 2018.
- [2] David R Dance and Ioannis Sechopoulos. Dosimetry in x-ray-based breast imaging. *Physics in Medicine & Biology*, 61(19):R271, 2016.
- [3] Etta D Pisano, R Edward Hendrick, Martin J Yaffe, Janet K Baum, Suddhasatta Acharyya, Jean B Cormack, Lucy A Hanna, Emily F Conant, Laurie L Fajardo, Lawrence W Bassett, et al. Diagnostic accuracy of digital versus film mammography: exploratory analysis of selected population subgroups in dmist. *Radiology*, 246(2):376–383, 2008.
- [4] R L Birdwell. The preponderance of evidence supports computer-aided detection for screening mammography. *Radiology*, 253(1):9–16, 2009.
- [5] N Karssemeijer, A M Bluekens, D Beijerinck, J J Deurenberg, M Beekman, R Visser, R van Engen, A Bartels-Kortland, and Mireille J Broeders. Breast cancer screening results 5 years after introduction of digital mammography in a population-based screening program. *Radiology*, 253(2):353–358, 2009.
- [6] L H Eadie, P Taylor, and A P Gibson. A systematic review of computer-assisted diagnosis in diagnostic cancer imaging. *European Journal of Radiology*, 81(1):e70 – e76, 2012.
- [7] Richard E Bird, Terry W Wallace, and Bonnie C Yankaskas. Analysis of cancers missed at screening mammography. *Radiology*, 184(3):613–617, 1992.
- [8] Roy JP Weber, Rob MG van Bommel, Marieke W Louwman, Joost Nederend, Adri C Voogd, Frits H Jansen, Vivianne CG Tjan-Heijnen, and Lucien EM Duijm. Characteristics and prognosis of interval cancers after biennial screen-film or full-field digital screening mammography. *Breast cancer research and treatment*, 158(3):471–483, 2016.
- [9] Junji Shiraishi, Qiang Li, Daniel Appelbaum, and Kunio Doi. Computer-aided diagnosis and artificial intelligence in clinical imaging. In *Seminars in nuclear medicine*, volume 41, pages 449–462. Elsevier, 2011.
- [10] Kunio Doi. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computerized medical imaging and graphics*, 31(4-5):198–211, 2007.

## BIBLIOGRAPHY

---

- [11] EB Cole, Z Zhang, HS Marques, RM Nishikawa, RE Hendrick, MJ Yaffe, W Padungchaichote, C Kuzmiak, J Chayakulkheeree, EF Conant, LL Fajardo, J Baum, C Gatsonis, and E Pisano. Assessing the stand-alone sensitivity of computer-aided detection with cancer cases from the digital mammographic imaging screening trial. *AJR Am J Roentgenol*, 199(3):392–401, 2012.
- [12] Constance D Lehman, Robert D Wellman, Diana SM Buist, Karla Kerlikowske, Anna NA Tosteson, and Diana L Miglioretti. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA internal medicine*, 175(11):1828–1837, 2015.
- [13] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.
- [14] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248, 2017.
- [15] Kenji Suzuki. Overview of deep learning in medical imaging. *Radiological physics and technology*, 10(3):257–273, 2017.
- [16] Thijs Kooi, Geert Litjens, Bram Van Ginneken, Albert Gubern-Mérida, Clara I Sánchez, Ritse Mann, Ard den Heeten, and Nico Karssemeijer. Large scale deep learning for computer aided detection of mammographic lesions. *Medical image analysis*, 35:303–312, 2017.
- [17] William Lotter, Greg Sorensen, and David Cox. A multi-scale cnn and curriculum learning strategy for mammogram classification. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 169–177. Springer, 2017.
- [18] L Tabár, P B Dean, and T Tot. *Teaching Atlas of Mammography*. Thieme, 3rd edition, 2001.
- [19] D T Ramsay, J C Kent, R A Hartmann, and P E Hartmann. Anatomy of the lactating human breast redefined with ultrasound imaging. *Journal of Anatomy*, 206(6):525–534, 2005.
- [20] R Smithuis and R Pijnappel. Breast - calcifications differential diagnosis, 2013.
- [21] breastcancer.org. Types of breast cancer. <http://www.breastcancer.org>, 2013.
- [22] W J H Veldkamp. *Computer Aided Characterization of Microcalcification Clusters in Mammograms*. PhD thesis, Radboud University Nijmegen Medical Center (The Netherlands), 2000.
- [23] M Samulski, R Hupse, C Boetes, R D M Mus, G J Heeten, and N Karssemeijer. Using computer-aided detection in mammography as a decision support. *European Radiology*, 20(10):2323–2330, 2010.

- [24] U Bick and F Diekmann. *Digital Mammography*. Springer, 3rd edition, 2010.
- [25] R Hupse. *Detection of malignant masses in breast cancer screening by computer assisted decision making*. PhD thesis, Radboud University Nijmegen Medical Center (The Netherlands), and Advanced School for Computing and Imaging (ASCI) graduate school, 2012.
- [26] H D Cheng, X Cai, X Chen, L Hu, and X Lou. Computer-aided detection and classification of microcalcifications in mammograms: a survey. *Pattern Recognition*, 36(12):2967 – 2991, 2003.
- [27] Timothy J Carroll. Trends in on-call workload in an academic medical center radiology department 1998–20021. *Academic radiology*, 10(11):1312–1320, 2003.
- [28] David C Levin, Vijay M Rao, Laurence Parker, and Andrea J Frangos. Analysis of radiologists’ imaging workload trends by place of service. *Journal of the American College of Radiology*, 10(10):760–763, 2013.
- [29] Carl D’Orsi, L Bassett, S Feig, et al. Breast imaging reporting and data system (bi-rads). *Breast imaging atlas, 4th edn. American College of Radiology, Reston*, 2018.
- [30] Joann G Elmore, Sara L Jackson, Linn Abraham, Diana L Miglioretti, Patricia A Carney, Berta M Geller, Bonnie C Yankaskas, Karla Kerlikowske, Tracy Onega, Robert D Rosenberg, et al. Variability in interpretive performance at screening mammography and radiologists? characteristics associated with accuracy. *Radiology*, 253(3):641–651, 2009.
- [31] Solveig Hofvind, Berta M Geller, Robert D Rosenberg, and Per Skaane. Screening-detected breast cancers: discordant independent double reading in a population-based screening program. *Radiology*, 253(3):652–660, 2009.
- [32] Robin N Strickland and Hee Il Hahn. Wavelet transforms for detecting microcalcifications in mammograms. *IEEE Transactions on Medical Imaging*, 15(2):218–229, 1996.
- [33] Issam El-Naqa, Yongyi Yang, Miles N Wernick, Nikolas P Galatsanos, and Robert M Nishikawa. A support vector machine approach for detection of microcalcifications. *IEEE transactions on medical imaging*, 21(12):1552–1563, 2002.
- [34] Athanasios Papadopoulos, Dimitrios I. Fotiadis, and Aristidis Likas. An automatic microcalcification detection system based on a hybrid neural network classifier. *Artificial intelligence in Medicine*, 25(2):149–167, 2002.
- [35] Jinshan Tang, Rangaraj M Rangayyan, Jun Xu, Issam El Naqa, and Yongyi Yang. Computer-aided detection and diagnosis of breast cancer with mammography: recent advances. *IEEE Transactions on Information Technology in Biomedicine*, 13(2):236–251, 2009.

## BIBLIOGRAPHY

---

- [36] Luis Claudio de Oliveira Silva, Allan Kardec Barros, and Marcus Vinicius Lopes. Detecting masses in dense breast using independent component analysis. *Artificial intelligence in medicine*, 80:29–38, 2017.
- [37] Ciro D’Elia, Claudio Marrocco, Mario Molinara, and Francesco Tortorella. Detection of clusters of microcalcifications in mammograms: A multi classifier approach. In *Int. Symp. on Computer-Based Med. Syst.*, pages 572–577. IEEE, 2008.
- [38] Claudio Marrocco, Mario Molinara, Ciro D’Elia, and Francesco Tortorella. A computer-aided detection system for clustered microcalcifications. *Artificial intelligence in medicine*, 50(1):23–32, 2010.
- [39] Afsaneh Jalalian, Syamsiah BT Mashohor, Hajjah Rozi Mahmud, M Iqbal B Saripan, Abdul Rahman B Ramli, and Babak Karasfi. Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review. *Clinical imaging*, 37(3):420–426, 2013.
- [40] Jan-Jurre Mordang, Tim Janssen, Alessandro Bria, Thijs Kooi, Albert Gubern-Mérida, and Nico Karssemeijer. Automatic microcalcification detection in multi-vendor mammography using convolutional neural networks. In *International Workshop on Digital Mammography*, pages 35–42. Springer, 2016.
- [41] Frank W Samuelson and Nicholas Petrick. Comparing image detection algorithms using resampling. In *3rd IEEE International Symposium on Biomedical Imaging: Nano to Macro, 2006.*, pages 1312–1315. IEEE, 2006.
- [42] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
- [43] Yoshua Bengio et al. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.
- [44] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.
- [45] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
- [46] Kevin Jarrett, Koray Kavukcuoglu, Yann LeCun, et al. What is the best multi-stage architecture for object recognition? In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2146–2153. IEEE, 2009.
- [47] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proc. of Int. Conf. Art. Int. and Stat.*, pages 249–256, 2010.
- [48] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

- 
- [49] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [50] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [51] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [52] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
- [53] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.
- [54] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [55] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- [56] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.
- [57] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.
- [58] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
- [59] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [60] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

## BIBLIOGRAPHY

---

- [61] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016.
- [62] L Wei, Y Yang, R M Nishikawa, M N Wernick, and A Edwards. Relevance vector machine for automatic detection of clustered microcalcifications. *IEEE Trans. Med. Imaging*, 24(10):1278–1285, 2005.
- [63] Ciro D’Elia, Claudio Marrocco, Mario Molinara, and Francesco Tortorella. Detection of clusters of microcalcifications in mammograms: A multi classifier approach. In *Computer-Based Medical Systems, 2008. CBMS’08. 21st IEEE International Symposium on*, pages 572–577. IEEE, 2008.
- [64] X Zhang, X Gao, and M Wang. MCs detection approach using bagging and boosting based twin support vector machine. In *IEEE Int. Conf. on Systems, Man and Cybernetics*, pages 5000–5505, 2009.
- [65] A Oliver, A Torrent, M Tortajada, M P Lladó, L Tortajada, M Sentís, and J Freixenet. A boosting based approach for automatic microcalcification detection. *Proceedings of International Workshop on Digital Mammography, LNCS*, 6136:251–258, 2010.
- [66] C Marrocco, M Molinara, C D’Elia, and F Tortorella. A computer-aided detection system for clustered microcalcifications. *Artificial Intelligence in Medicine*, 50(1):23 – 32, 2010.
- [67] I El Naqa, Wernick M N Yang, Y, N P Galatsanos, and R M Nishikawa. A support vector machine approach for detection of microcalcifications. *IEEE Trans. Med. Imaging*, 21(12):1552–1563, 2002.
- [68] A. Bria, N. Karssemeijer, and F. Tortorella. Learning from unbalanced data: A cascade-based approach for detecting clustered microcalcifications. *Medical Image Analysis*, 18(2):241–252, 2014.
- [69] P Viola and M Jones. Robust real-time object detection. *Int. J. of Comp. Vis.*, 57(2):137–154, 2001.
- [70] Alessandro Bria, Claudio Marrocco, Nico Karssemeijer, Mario Molinara, and Francesco Tortorella. Deep cascade classifiers to detect clusters of microcalcifications. In *International Workshop on Digital Mammography*, pages 415–422. Springer, 2016.
- [71] Alessandro Bria, Claudio Marrocco, Mario Molinara, Benedetta Savelli, Jan-Jurre Mordang, Nico Karssemeijer, and Francesco Tortorella. Improving the automated detection of calcifications by combining deep cascades and deep convolutional nets. In *14th International Workshop on Breast Imaging, IWBI 2018, Atlanta, Georgia, USA, 8-11 July 2018*, page 1071808, 2018.
- [72] Alessandro Bria, Claudio Marrocco, Nico Karssemeijer, Mario Molinara, and Francesco Tortorella. Deep cascade classifiers to detect clusters of microcalcifications. In *International Workshop on Breast Imaging*, pages 415–422. Springer, 2016.

- [73] Jan-Jurre Mordang, Tim Janssen, Alessandro Bria, Thijs Kooi, Albert Gubern-Mérida, and Nico Karssemeijer. Automatic microcalcification detection in multi-vendor mammography using convolutional neural networks. In *International Workshop on Digital Mammography*, pages 35–42. Springer, 2016.
- [74] Alessandro Bria, Claudio Marrocco, Jan-Jurre Mordang, Nico Karssemeijer, Mario Molinara, and Francesco Tortorella. Lut-qne: Look-up-table quantum noise equalization in digital mammograms. In *International Workshop on Digital Mammography*, pages 27–34. Springer, 2016.
- [75] F W Samuelson and N Petrick. Comparing image detection algorithms using resampling. In *IEEE Int. Symp. Biomed. Imag.*, pages 1312–1315, 2006.
- [76] O J Dunn. Multiple Comparisons Among Means. *Journal of the American Statistical Association*, 56(293):52–64, 1961.
- [77] G. Bradski. The OpenCV Library. *Dr. Dobb’s Journal of Software Tools*, 2000.
- [78] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv preprint arXiv:1408.5093*, 2014.
- [79] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proc. of the IEEE Int. Conf. on Computer Vision*, pages 1026–1034, 2015.
- [80] Jose Bernal, Kaisar Kushibar, Daniel S Asfaw, Sergi Valverde, Arnau Oliver, Robert Martí, and Xavier Lladó. Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review. *Artificial intelligence in medicine*, 95(3):64–81, 2018.
- [81] Damla Arifoglu and Abdelhamid Bouchachia. Detection of abnormal behaviour for dementia sufferers using convolutional neural networks. *Artificial intelligence in medicine*, 94:88–95, 2019.
- [82] Ghulam Murtaza, Liyana Shuib, Ainuddin Wahid Abdul Wahab, Ghulam Mujtaba, Henry Friday Nweke, Mohammed Ali Al-garadi, Fariha Zulfiqar, Ghulam Raza, and Nor Aniza Azmi. Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges. *Artificial Intelligence Review*, pages 1–66, 2019.
- [83] Zhenghao Shi, Huan Hao, Minghua Zhao, Yaning Feng, Lifeng He, Yinghui Wang, and Kenji Suzuki. A deep CNN based transfer learning method for false positive reduction. *Multimedia Tools and Applications*, pages 1–17, 2018.

## BIBLIOGRAPHY

---

- [84] Chaofeng Li, Guoce Zhu, Xiaojun Wu, and Yuanquan Wang. False-positive reduction on lung nodules detection in chest radiographs by ensemble of convolutional neural networks. *IEEE Access*, 6:16060–16067, 2018.
- [85] Ludmila I Kuncheva. *Combining pattern classifiers: methods and algorithms, 2nd ed.* John Wiley & Sons, 2014.
- [86] Claudio Marrocco, Mario Molinara, and Francesco Tortorella. Exploiting AUC for optimal linear combinations of dichotomizers. *Pattern Recognition Letters*, 27(8):900–907, 2006.
- [87] Maria Teresa Ricamato, Claudio Marrocco, and Francesco Tortorella. MCS-based balancing techniques for skewed classes: An empirical comparison. In *19th Int. Conf. on Patt. Rec.*, pages 1–4. IEEE, 2008.
- [88] C. De Stefano, F. Fontanella, C. Marrocco, and A. Scotto di Freca. A hybrid evolutionary algorithm for bayesian networks learning: An application to classifier combination. In *PART 1 LNCS*, volume 6024, pages 221–230. Springer, 2010.
- [89] C. De Stefano, G. Folino, F. Fontanella, and A. Scotto Di Freca. Using bayesian networks for selecting classifiers in GP ensembles. *Information Sciences*, 258:200–216, 2014.
- [90] Hongwei Li, Gongfa Jiang, Jianguo Zhang, Ruixuan Wang, Zhaolei Wang, Wei-Shi Zheng, and Bjoern Menze. Fully convolutional network ensembles for white matter hyperintensities segmentation in mr images. *NeuroImage*, 183:650 – 665, 2018.
- [91] Ariel Benou, Ronel Veksler, Alon Friedman, and T Riklin Raviv. Ensemble of expert deep neural networks for spatio-temporal denoising of contrast-enhanced mri sequences. *Medical image analysis*, 42:145–159, 2017.
- [92] Julian Zilly, Joachim M Buhmann, and Dwarikanath Mahapatra. Glaucoma detection using entropy sampling and ensemble learning for automatic optic cup and disc segmentation. *Computerized Medical Imaging and Graphics*, 55:28–41, 2017.
- [93] Benedetta Savelli, Claudio Marrocco, Alessandro Bria, Mario Molinara, and Francesco Tortorella. Combining convolutional neural networks for multi-context microcalcification detection in mammograms. In *Computer Analysis of Images and Patterns - CAIP 2019 International Workshops, ViMaBi and DL-UAV, Salerno, Italy, September 6, 2019, Proceedings*, pages 36–44, 2019.
- [94] Mario Molinara Claudio Marrocco Francesco Tortorella Benedetta Savelli, Alessandro Bria. A multi-context cnn ensemble for small lesion detection. *Artificial Intelligence in Medicine*, In press.
- [95] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31, 2017.



- [96] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.
- [97] Juan Wang and Yongyi Yang. A context-sensitive deep learning approach for microcalcification detection in mammograms. *Pattern recognition*, 78:12–22, 2018.
- [98] Inês C Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria João Cardoso, and Jaime S Cardoso. Inbreast: toward a full-field digital mammographic database. *Academic radiology*, 19(2):236–248, 2012.
- [99] Yann A LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer, 2012.
- [100] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, et al. Caffe: Convolutional architecture for fast feature embedding. In *Proc. of the 22nd Int. Conf. on Multimedia*, pages 675–678. ACM, 2014.
- [101] Alessandro Bria, Claudio Marrocco, Mario Molinara, and Francesco Tortorella. An effective learning strategy for cascaded object detection. *Information Sciences*, 340-341:17 – 26, 2016.
- [102] Meindert Niemeijer, Bram Van Ginneken, Michael J Cree, et al. Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs. *IEEE Trans. on Medical Imaging*, 29(1):185–195, 2009.
- [103] Alessandro Bria, Nico Karssemeijer, and Francesco Tortorella. Learning from unbalanced data: a cascade-based approach for detecting clustered microcalcifications. *Medical image analysis*, 18(2):241–252, 2014.
- [104] Liyang Wei, Yongyi Yang, Robert M Nishikawa, and Yulei Jiang. A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications. *IEEE transactions on medical imaging*, 24(3):371–380, 2005.
- [105] Juan Wang, Robert M Nishikawa, and Yongyi Yang. Global detection approach for clustered microcalcifications in mammograms using a deep learning network. *Journal of Medical Imaging*, 4(2):024501, 2017.
- [106] Rui Hou, Yinhao Ren, Lars J Grimm, Maciej A Mazurowski, Jeffrey R Marks, Lorraine King, Carlo C Maley, E Shelley Hwang, and Joseph Y Lo. Malignant microcalcification clusters detection using unsupervised deep autoencoders. In *Medical Imaging 2019: Computer-Aided Diagnosis*, volume 10950, page 109502Q. International Society for Optics and Photonics, 2019.

## BIBLIOGRAPHY

---

- [107] Heng-Da Cheng, Xiaopeng Cai, Xiaowei Chen, Liming Hu, and Xueling Lou. Computer-aided detection and classification of microcalcifications in mammograms: a survey. *Pattern recognition*, 36(12):2967–2991, 2003.
- [108] Yu Zhang and Qiang Yang. A survey on multi-task learning. *arXiv preprint arXiv:1707.08114*, 2017.
- [109] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.
- [110] Jie Xu, Lin Zhao, Shanshan Zhang, Chen Gong, and Jian Yang. Multi-task learning for object keypoints detection and classification. *Pattern Recognition Letters*, 2018.
- [111] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):121–135, Jan 2019.
- [112] Mohammed A Al-Masni, Mugahed A Al-Antari, Jeong-Min Park, Geon Gi, Tae-Yeon Kim, Patricio Rivera, Edwin Valarezo, Mun-Taek Choi, Seung-Moo Han, and Tae-Seong Kim. Simultaneous detection and classification of breast masses in digital mammograms via a deep learning yolo-based cad system. *Computer methods and programs in biomedicine*, 157:85–94, 2018.
- [113] Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, Jul 1997.
- [114] Jonathan Baxter. A bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine Learning*, 28(1):7–39, Jul 1997.
- [115] Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 845–850, 2015.
- [116] Yongxin Yang and Timothy M Hospedales. Trace norm regularised deep multi-task learning. *arXiv preprint arXiv:1606.04038*, 2016.
- [117] Yulei Jiang, Robert M Nishikawa, and John Papaioannou. Dependence of computer classification of clustered microcalcifications on the correct detection of microcalcifications. *Medical Physics*, 28(9):1949–1957, 2001.
- [118] Juan Wang and Yongyi Yang. Spatial density modeling for discriminating between benign and malignant microcalcification lesions. In *2013 IEEE 10th International Symposium on Biomedical Imaging*, pages 133–136. IEEE, 2013.

- [119] María V Sainz de Cea, Robert M Nishikawa, and Yongyi Yang. Estimating the accuracy level among individual detections in clustered microcalcifications. *IEEE transactions on medical imaging*, 36(5):1162–1171, 2017.
- [120] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
- [121] Xiangyang Xu, Shengzhou Xu, Lianghai Jin, and Enmin Song. Characteristic analysis of otsu threshold and its applications. *Pattern recognition letters*, 32(7):956–961, 2011.
- [122] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
- [123] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

## BIBLIOGRAPHY

---