Computer aided diagnosis of breast cancer in mammography using deep neural networks

Thijs Kooi

Contents

Chapter 1

Introduction

1.1 Artificial intelligence

Archaeologists have found evidence that our ancestors started employing tools to manipulate their environment as early as 2.6 million years ago. Hammers allowed them to open nuts, spears to hunt and fire to process food more efficiently. The most expedient tool wielders increased their fitness and secured an evolutionary advantage over their less sapient counterparts. As we and our brains evolved, so did these tools, into what is most likely in your pocked or in front of you right now: the microchip. Where early implements such as hammers and spears were better substitutes for our hands and arms, computers are increasingly shown to be better substitutes for our brains and have proven themselves indispensable in landmark intellectual achievements such as the moon landing or the large hadron collider.

Early computers were particularly good at cognitive tasks humans were particularly bad at, like adding large numbers and storing data consistently, but failed miserably at tasks humans ostensibly perform effortlessly, such as processing visual or auditorial stimuli (often tasks that require approximation algorithms to be solved in reasonable time, known as NP-hard problems). This phenomenon was first verbalized by Hans Moravec [189] and became known as the Moravec paradox. The scientific field of *Artificial Intelligence* is dedicated to the creation of computer programs that turn Moravec's paradox around and make systems that perform speech and language processing, vision and reasoning tasks at (super)human level.

As one of the few scientific fields to have an official birthday, AI traces its Dies Natalis to the Dartmouth conference of August 1956. Most participants there were young pioneers in the field at the time. At this event it was conjectured that artificial general intelligence (AGI), a system performing on par to humans in all aspects of cognition, would be possible in no more than a few decades after the conference took place, which proved to be a gross underestimate. This type of optimism and failure to deliver characterized the life cycle of the field and led to several 'AI summers' and 'AI winters' where audacious claims spurred by brief progress were followed by cuts in funding when projects failed to live up to expectations.

Throughout its development as a scientific discipline, AI saw a strong divide between the *connectionist* camp that borrowed techniques from cognitive neuroscience, statistics and theoretical physics and the *symbolic* camp, that relied on rule based systems with foundations in formal logic. Most early success in AI was achieved by the latter, in particular in the form of expert systems: software that can reason about a large preprogrammed knowledge base using deterministic rules. These systems were applied successfully to diagnose disease or errors in machines, planning and monitoring. Other narrow and deterministic fields such as games also proved particularly accommodating to symbolic AI, most notably the watershed victory of Deep Blue, a chess computer, over world champion Gary Kasparov on 10 February 1996.

At the core of logic based systems were algorithms that sought out steps towards an optimal solution. Before the theory of computational complexity was mature enough, it was falsely assumed that for larger problems, the same procedures could just be followed, only with better hardware [221] and Go, a board game with far more possible moves than chess at each point in the game, would be solved in a similar fashion. Additionally, writing programs based on logic and rules for less clear-cut tasks, such as speech and vision proved to be more difficult then initially anticipated and developed systems were brittle and broke down completely when presented with input not accounted for during the development of

the algorithm.

By the mid 1990s, the limitations of symbolic AI became increasingly clear and nature inspired algorithms such as neural networks and evolutionary computation took center stage. These models rely less on preprogrammed rules based on knowledge of experts and instead use techniques from mathematical statistics and computational neuroscience to learn about a problem based on examples and to come up with an approximate solution that fits these examples as well as possible. Although neural networks, models inspired by how the brain processes information, were an important branch, numerous algorithms with varying degrees of biological plausibly have been proposed. Nevertheless, given the analogy to human learning, the field became known as machine learning.

1.1.1 Machine learning

Machine learning (ML) algorithms are used to perform tasks that we do not know how to program or are difficult to program, by providing a predefined model of the task with training data. This model has some input, output and is equipped with dials that can be tuned to predict the output of an input for every sample in a 'training set' as well as possible, a type of learning referred to as *supervised learning*. When the output is an instance of a predefined set of classes, the task is referred to as a *classification* problem, if the output is continuous, it is known as a *regression* problem. When no correct output is provided the process is called *unsupervised learning*, which is often used for finding structure in training data, such as groups that have some common traits.

Machine learning algorithms are already used in abundance, even though most people will not notice them: Google learns from results you and other users click on, Amazon, Netflix and Yelp will learn about your preferences, your bank and credit card company have fraud detection algorithms in place and your phone, smart watch and fitness trackers are full of methods that detect your activity, predict your directions and many more behavioral traits [65]. More conspicuous areas, mostly due to mistakes these systems make, are still speech recognition and synthesis, machine translation and computer vision applications such as face recognition, object recognition and self-driving vehicles.

The amount of tunable dials in the model- the number of parameters, is often related to the size of an example in the training data. When learning tasks such as fraud detection, an example can be a list of properties of a person. For more complex problems such as vision and speech recognition, one example is typically an image or audio signal. If the raw signal is given to the model without further preprocessing, a large amount of parameters are needed, which makes the problem difficult to learn and requires massive amounts of annotated data to make reasonable estimates of good values for these parameters. Consequently, traditional machine learning systems operating on sound or image signals typically relied on *features*: elements of the signal engineers or domain experts think are most descriptive for the classification problem, yet ignore all unnecessary factors. When doing face recognition, one can think of features such as shape of the eyes, mouth and nose. For medical problems doctors or other experts are often consulted. Examples of features used for tumor detection are the shape of a potential lesion, the contrast to its surroundings and textural properties. By far, most of the engineering in feature based systems is spent on developing these summarizations of the signal.

Feature extraction provides a platform to instill task-specific, a-priori knowledge that can be difficult to learn from data, but also causes a large bias towards how we humans think the task is performed, which may not be the optimal strategy. Firstly, because how we solve the task may simply not be optimal. Evolution gets stuck in local optima: land animals do not have wheels, even though these are currently the most efficient way to propel ourselves. Secondly, developing features by introspection: thinking how we think the task is performed in our brain, is difficult. Studies show characteristics such as attractiveness of a face are determined in a fraction of a second [288], yet if you ask a person, few if any will be able to provide an exhaustive list of what type of features they used to base their judgment on.

Since the inception of AI as a scientific discipline, research has seen a shift from rule-based, problem specific solutions to increasingly generic, problem agnostic methods that rely on training data. For many image and speech analysis tasks, more generic features started to emerge in the 2000s, such as Mel-frequency cepstral coefficients (MFCC)s [177] for audio and histograms of gradients such as SIFT [179] for images. *Representation learning* takes this idea one step further: engineers define a functional form with some parameters of a 'feature extractor' and the optimal parameters are learned based on data that best summarize the signal, by looking at dominant sources of variation in the data, typically ignoring the labels initially.

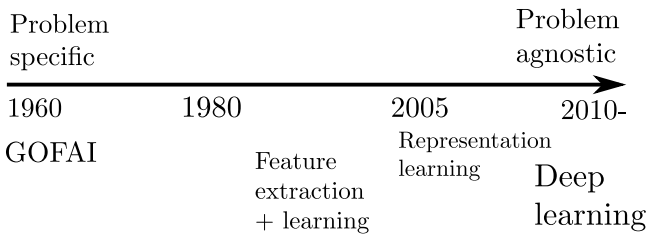


Figure 1.1: Illustration of the development of AI and machine learning based systems. Early systems (GOFAI = good old fashioned AI) relied largely on engineers and many handcrafted rules to make them work, while contemporary systems rely on large amounts of data and less on domain experts.

Although representation learning had some success, a problem with this approach is that the dominant sources of variation may not correspond to parts of the data that are useful to predict the output as well as possible. For example, when doing face detection, the intensity of images can vary a lot between photos, but is not relevant to identify the person. *Deep learning* [12, 10, 229, 163] is the latest and some would argue final stage in the shift towards generic AI systems and describes models learned end-to-end from data. Figure 1.1 provides a schematic overview of the AI timeline.

1.1.2 Deep learning

Since 2010, machine learning went through a Cambrian explosion with the introduction of methods to train deep neural networks: statistical models stacked in such a way that it (almost magically) allows them to crack problems which half a decade ago seemed insurmountably complex. The idea is not new and work has been done since the late seventies [90, 162]. Compounded by a lack of large curated datasets, computational power and efficient training methods, early 'deep' machine learning algorithms failed to live up to the inventor's expectations. In 2006, however, two papers [119, 11] emerged that showed deep networks can be trained by stacking models and training these individually, without labels. The resulting system was then fine tuned in its entirety using the labels in the dataset. Currently, the most efficient models are trained end-to end, without complicated stage-wise training procedures.

A major catalyst behind the success of deep neural networks was the availability of large datasets such as ImageNet [56], an enormous database of nearly 15 million manually labeled natural images, organized into synonym sets (classes of objects or other elements in the image) each consistent of about 1000 images. These datasets provided the copious parameters in the models with enough examples to make reasonable estimates of their optimal value. Until 2012, feature based systems were used and error rates of around 30 percent were claiming top positions. When the first deep learning based method was introduced in 2012, the error rate dropped to 4% in a matter of three years.

Although at the moment, large datasets are necessary for good performance, their size imposes an additional burden on the hardware required to train the models. Traditionally, neural networks were trained by adjusting the parameters that best - on average - predict the training set. ImageNet or other comparably large datasets, do not fit in the memory of the typical desktop computer and commodity hardware is not powerful enough to perform training in reasonable time. Instead, the parameters are fit by taking small subsets of the data and processing these batches repeatedly on a graphical processing unit (GPUs), a type of processor specialized in parallel computation that were originally designed for gaming. Currently, specialized hardware such as Google's tensor processing unit (TPU) is being developed and used.

Deep learning algorithms have already well penetrated into products and major tech companies like Google, Facebook, Microsoft and Baidu are assigning billion dollar R&D budgets to the development and deployment of these methods. Even companies that do not have software or AI as their main focus like Uber, Amazon, Twitter, Netflix, AirBnB, etc. are following suit and investing heavily. Due to the current supply-demand ratio, companies are paying hefty premiums for top experts in the field and six or even seven figure salaries are not unheard of. Although academic institutions are grudgingly seeing their scientist leave in favor of industry, funding from industry is currently abundant and efficient collaborations between academia and industry are starting to emerge. The deep learning hype launched us into an AI summer that is hotter than ever.

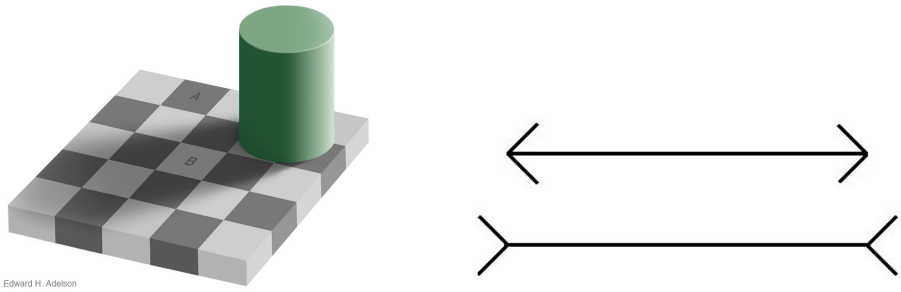


Figure 1.2: Humans are poor at observing quantitative information, both in intensity (a) and shape (b).
(a) Adelson's checker shadow illusion (b) Muller Lyer illusion.

1.2 AI in medicine

Just like our ancestors developed tools to crack open nuts or hunt animals to increase their fitness and chances of survival, so can we employ our contemporary tools to perform tasks that increase our life expectancy. In that spirit, medicine has been receiving a lot of attention from engineers, initially in the form of hardware such as imaging techniques, but recently also in form of software and machine learning based systems. Computers are consistent, not subject to fatigue and have the potential to learn from immense amounts of data, far more than any medical expert will experience in her or his lifetime.

Additionally, there are currently several limitations to humans that computers do not suffer from. Our visual faculties have been shaped by millions or years of evolution, which made them relatively efficient for processing the world around us. For most of the time however, this did not involve reading medical images and finding potential abnormalities and therefore, knowing the exact wavelength of our input data was not important: a tiger is dangerous during both sunset and at noon. To a lesser extent the same principles apply to observing size and shape. Objects are often evaluated based on the scene around it, something designers of mad houses make eager use of to fool us. This poor quantitative vision is best illustrated by the two famous examples in figure 1.2.

More so than vision, computers can and already are outperforming humans in reasoning tasks. As early as 1976 [29], the MYCIN system [238], an expert system used to classify gut bacteria, outperformed human diagnostics. Daniel Kahnemann and colleagues [272] identified numerous cognitive biases and show where humans make suboptimal choices. These had a particular application in economics, but are also important in medicine. An example of such a bias is that humans consistently overestimate the likelihood of rare diseases [53, 17, 22], causing overdiagnosis, resulting in unnecessary stress for patients and economic burden for clinics.

These shortcomings in human cognitive capacities can have even more serious consequences than overdiagnosis. Human error in medicine has recently been identified as the third leading cause of death in the US [1] [181]. However, it also provides great opportunities for computer scientists and AI specialists to improve medical care. *Computer aided diagnosis* (CAD) [99, 64, 63, 277] is a field dedicated to the development of AI systems to support and ultimately replace diagnostic tasks of medical experts. Although CAD could in principle be applied to any type of medical data, at the moment it is still mostly synonymous with the automatic interpretation of medical images. Two types of CAD are sometimes distinguished: computer aided detection (CADe) and computer aided diagnosis (CADx). In the first setting, the computer simply pinpoints an abnormality and in the second setting the computer is used for interpretation. From a machine learning perspective, however, these are equivalent and I will simply use CAD throughout this thesis to refer to both.

With increasing life span, cancer has become a major health problem and is now the leading cause of death in the US [242]. After lung cancer, prostate cancer and *breast cancer* are the most lethal cancers for men and women, respectively. Since efforts to effectively stop humans from aging are still in their infancy, working on techniques to prevent and cure these diseases is an excellent path towards increasing our fitness and life expectancy.

[1] http://www.forbes.com/sites/leahbinder/2013/09/23/stunning-news-on-preventable-deaths-in-hospitals/

1.3 Breast cancer

Cancer is an umbrella term for a myriad of diseases related to uncontrolled cell growth. Colloquially, the terms tumor and cancer are often used interchangeably. However, not all cancers cause tumors and not all tumors are cancer; a tumor refers to a mass of cells, which can be benign or malignant. The main types of cancer are *carcinoma*, *sarcoma*, *melanoma*, *lymphoma* and *leukemia*, classified by their tissue of origin. The vast majority of cancers are carcinoma, originating from epithelial tissue. Two types of genes are commonly held responsible for the formation of cancer: (1) *oncogenes* and (2) *tumor suppressor genes*. Oncogenes, when functioning normally, cause cells to grow and divide. When mutations occur and they malfunction, cell growth is out of control. Tumor suppressor genes on the other hand can slow down cell division and control when cells should die (apoptosis). Mutations on any of these genes can significantly increase an individual's risk to develop cancer.

As mentioned above, breast cancer is one of the most common cancers. It is estimated that in 2017 in the US, around 250000 women will be diagnosed with invasive breast cancer and another 60000 with an early stage or non-invasive variant [242]. Breast cancer is currently the most prevalent cancer among women in the US, accounting for 29% of new cancer diagnosis and is responsible for 14% of all cancer deaths, making it also the second most lethal cancer in the female population. One in eight women will develop invasive breast cancer somewhere in their lifetime [242]. Research on breast cancer has made significant advances in the past decades, with five year survival rates in the US now up to 90% over the period 2003-2009 [58, 222]. In spite of this, it remains one of the most lethal cancers in absolute numbers with an estimated 40000 deaths annually in the US [242], 11000 in the UK and about 3000 in the Netherlands.

Breast cancer types

Most breast cancers are *carcinoma*, meaning they originate from epithelial tissue and form in parts of the breast comprised of lobules and ducts, responsible for milk production. An illustration is provided in figure 1.3. These carcinoma are again split up into four types, depending on their place of origin and stage of proliferation.

- **Invasive Ductal Carcinoma** (IDC): An invasive or infiltrating cancer originating from the milk ducts. This is the most common form of breast cancer accounting for roughly 85% of all cases.
- **Invasive Lobular Carcinoma** (ILC): Originates from milk producing glands (lobules) and accounts for about 15% of invasive cancers. It is often harder to spot on images such as mammograms.
- **Ductal carcinoma in situ** (DCIS): DCIS comes in low, medium and high grade. The current way of working is to remove them all because of a substantial risk of evolving into high risk invasive tumors. Confronted with this diagnosis most women do not take any chances and have it removed. However this is partly caused by the name carcinoma, and a considerable number of medium and low grade DCIS would never have been diagnosed without screening. Some people call this overdiagnsosis and is often brought up as the downside of screening.
- **Lobular Carcinoma In Situ** (LCIS): A rare form of in situ breast cancer occurring in the lobules. It is generally not considered malignant or a pre-cursor to a malignant tumor, but has been associated with an increased risk of developing a malignant mass, meaning closer surveillance but no treatment as such.

If untreated, breast cancer can spread to the rest of the body through the lymphatic system and can be lethal. When a patient or doctor suspects there is an anomaly, images are often recorded.

1.3.1 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, or sometimes referred to as a type of mammography, tomosynthesis is based on X-ray, but instead a series of X-rays is taken and a reconstruction is made that allows the reader to scroll through the breast in slices and get a better view of structures that would otherwise be hidden. The dose of radiation

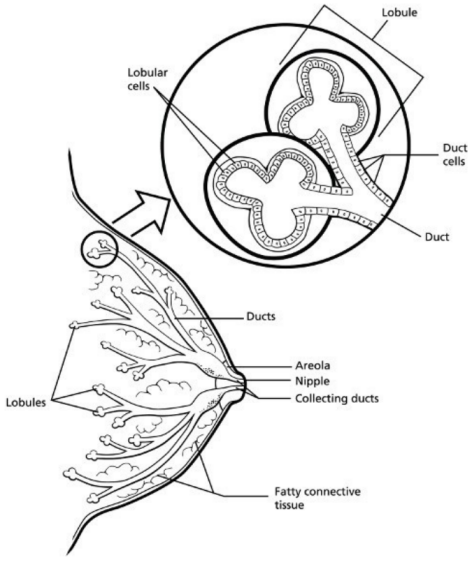


Figure 1.3: Illustration of breast tissue (image taken from cancer.org)

is slightly higher though within the limits of safe radiation outlined by the FDA. Tomosynthesis has been shown, depending on the vendor and the way of working, to be equal or improve the detection of breast cancer with fewer false positives [248, 247].

- **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 (ABUS)** 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 a 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. It is therefore unsuitable for stand-alone imaging [2].
- **Magnetic Resonance Imaging** (MRI) MRI uses magnetic fields and radiowaves to generate images of internal structure. Similar to ABUS, 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 [3] and in general with a lower specificity.

Since mammography is the subject of this thesis, I will go into slightly more detail about the process.

1.3.2 Mammography

Mammography is the oldest and still most common breast imaging technique. Mammograms, as the resulting recordings are called, are taken using dedicated X-ray machines. The breast is flattened between two plates and compressed. The compression spreads tissue, making it easier to find tumors and decreases the amount of exposure needed, though is often experienced as painful by patients. Initially, images were recored using an analogue system, known as screen film mammography (SFM) that printed an image onto a large sheet of film. Currently, digital mammography (DM), also sometimes referred to as full field digital mammography (FFDM) is most commonly used. DM has been shown to be at least as good as SFM, but has several other advantages. Since images are already in a digital form, they are easily stored and transmitted to other institutions and computer systems have access to raw values in the image, meaning less loss of information.

[2] https://www.radiologyinfo.org/en/info.cfm?pg=breastus
[3] https://www.radiologyinfo.org/en/info.cfm?pg=breastmr

Figure 1.4: Example of a mammographic exam with an obvious mass. The left two images are the MLO views and right two images the CC views.

As mentioned above, the amount of radiation that is absorbed by the detector depends on the attenuation properties of the tissue it passes through. If an x-ray photon initially has energy I_0 and passes through tissue with attenuation coefficient μ and height h, the resulting energy follows Beers law:

$$I = I_0 \exp\{-h\mu\} \tag{1.1}$$

The resulting energy in digital mammograms is usually stored in 12 or 14 bits and is typically inverted such that white corresponds to structures with high attenuation and vice versa. Because the pixel value is effectively a summation (after a log transform) the mapping is surjective: many different 3D structures can result in the same image, making potential malignancies difficult to see. Consequently, two images are typically recorded the craniocaudal (CC) and mediolateral oblique (MLO) views. The MLO view often has a large portion of the pectoral muscle in the image, which is also visible, when possible on the CC. Figure 1.4 shows an example of the four typical images obtained in an exam with a malignant mass in the right/left breast.

Mammography is used in both a diagnostic and screening setting. In the first case, for women who have symptoms such as a palpable lump or who have been referred from a screening and in the second setting for women at risk of developing breast cancer.

1.3.3 Breast cancer screening

Breast cancer screening in the form of annual or biennial breast X-rays is being performed to detect cancer at an early stage. In most western countries this has been implemented since the early 21st century. In the Netherlands screening started as early as 1988 [4]. Currently, 40 million mammograms are read in the US, roughly 2 million in the UK and 1 million in the Netherlands, on a yearly basis. Screening has been shown to significantly reduce breast cancer mortality by up to 40% [262, 27, 222] and early detection also allows for a greater range of less invasive treatments. A downside of this type of imaging is that the radiation fed through the tissue increases the risk of developing cancer. However, the benefits of screening have been shown to outweigh the risk with benefit-to-risk ratios ranging from 8:1 to 20:1 [130].

Unfortunately, cancers are occasionally missed during screening. Some are undetectable and simply not visible on the mammogram due to factors such as high breast density [37] or occlusion, but about 20% to 30% of tumors are missed in screening [290, 75] that were visible in hindsight. Evans et al. [76] showed that mammographers are more likely to miss tumors if they are in a low incidence setting, such as a screening, going from 30% false negatives to 12% when the same tumors are present in an equal incidence setting, indicating that drops in performance may be partially responsible for missed malignancies. Additionally, unnecessary referrals have been held responsible for overdiagnosis [20, 254]. For instance, Bleyer et al. [20] estimated that 1.3 million women or about 30% of all screened women in the US, were over-diagnosed in the past 30 years and received treatment for cancer that would never have led to clinical symptoms. Luckily,

[4] http://www.rivm.nl/Onderwerpen/M/Medische_Stralingstoepassingen/Trends_en_stand_van_zaken/Diagnostiek/Extramurale_radiologie/Mammografie_screening

other studies find much lower numbers [295] with rates as low as 10% for the ages 55-69, but still indicate potential for improvement.

The screening protocol varies widely per country and organization. In the Netherlands, UK and many other European countries, women are invited for screening from the age of 50, in the US and east Asia this is 40. Women in high risk populations like BRCA1 and BRCA2 carriers, are typically screened from a much younger age and receive additional imaging such as an MRI. In the Netherlands, from 1000 women, about 25 are referred for a second exam, typically called a recall. From those 25 women, about 18 turn out not to have cancer, meaning that 7 out of 1000 screened women actually have cancer. In the first screening round, this number is higher [5].

Similar to the screening protocol, the way in which images are read also varies widely per country. In some countries like the Netherlands, independent double reading is being performed: each image is read by two radiologists, who only communicate about the case if there is a disagreement in their label. In some countries or institutions, the images are read by a single radiologist or a trained nurse. To standardize interpretation of mammograms, the Breast Imaging Reporting and Database System (BI-RADS) system has been introduced. It includes 6 standardized categories, each with their associated follow-up plan.

1.3.4 Signs of breast cancer in mammography

The main signs of malignancy on mammography are:

- **Masses or soft tissue lesions** Malignant masses are the most common sign of breast cancer on mammography. The size and shape varies, but they are often characterized by high density and a contracting, spiculation pattern originating from its center.
- **Calcifications** Calcifications look like small white flecks and can indicate the presence of an early stage cancer or serve as an additional cue for the interpretation of a potentially invasive mass. Calcifications are often associated with DCIS and can therefore serve as a precursor for tumor.
- **Architectural distortion** Identified as the third most common sign of malignant breast cancer [95]. They often display a contracting pattern similar to masses, but without a clear density in their center.
- **Asymmetry** An asymmetry refers to a potentially malignant density that is not characterized as a mass or architectural distortion. Global asymmetries have been associated with an increased risk of developing cancer [232].
- **Lymph node** Lymph nodes are part of normal tissue, but can be confused with tumor. Enlarged lymph nodes in the center of the breast can indicate the spread of a malignant cancer and are particularly suspicious in combination with another site that resembles as mass or calcifications.

Some of the most common benign abnormalities are:

- **Cysts** Cysts are small fluid filled sacks and are common. It is estimated that 7% of women in the western world will develop palpable breast cysts in their lifetime. Even though cysts have been correlated with risk of developing breast cancer [62], many of them are benign and do not require follow-up. On mammography, benign cysts and solid lesion can be difficult to discriminate and consequently, many women are being recalled unnecessarily for a second diagnostic exam or core needle biopsy. Literature suggest 20% [72] to 37% [240] of recalls can be attributed to benign solitary cysts.
- **Radial scar** A radial scar is a form of sclerosing duct hyperplasia and can cause an architectural distortion. Radial scars may indicate a disturbance in the breast tissue and more specifically between the stromal (supportive) and functional elements (lobules, ducts, etc), that can lead to the formation of scar tissue, or possibly to cancer. [6]
- **Fibroadeanoma** Fibroadenomas are benign tumors made up of both glandular breast tissue and stromal (connective) tissue. The are most common in young women in their 20s and 30s. A woman can have one or many

[5] http://www.rivm.nl/Onderwerpen/B/Bevolkingsonderzoek_borstkanker/Veelgestelde_vragen/Uitslag_en_vervolgonderzoek

[6] http://breast-cancer.ca/radscrs/

fibroadenomas. Women with fibroadenomas have an increased risk of breast cancer – about 1 to 2 times the risk of women with no breast changes. Fibroadenomas are the most common kind of breast mass in young women.[7]

- **Breast Arterial Calcifications** (BACs) Arterial calcifications have been associated with cardiovascular disease [286], but can be confused with cancerous calcifications. Since BACs grow inside arteries, they typically exhibit a (curved) linear structure, contrary to malignant calcifications that have a more scattered pattern.

1.4 Computer aided diagnosis

CAD is being developed for a variety of pathologies and modalities but mammography has thus far been on the forefront of this development stemming from the early nineties [213, 182, 71, 197, 6]. In June 1998, the first CAD system for mammograms developed by R2 Technology Inc. was approved by the US Food and Drug Administration (FDA). In the US in 2010 about to 70% of all screening studies in hospital facilities and 85% in private institutions [213] employ CAD as an additional reader. Recent estimates indicate that 90% of mammograms in the US are read by CAD. Unfortunately, performance has mostly stagnated in the past decade. Methods are mostly being developed on small datasets [192, 298] which are not always shared and algorithms are difficult to compare [71].

As mentioned above, breast cancer has two main manifestations in mammography, firstly the presence of malignant soft tissue or masses and secondly the presence of microcalcifications [42]. Different systems are currently being developed for masses and calcifications, although given recent developments in machine learning, this may soon change. Microcalcifications are often small and can easily be missed by oversight. Some studies suggest CAD for microcalcifications is highly effective in reducing oversight [182] with acceptable numbers of false positives.

Unfortunately, the merit of CAD for masses is less clear. Research suggests human errors do not stem from oversight but rather misinterpretation [182] and some studies show no increase in sensitivity or specificity with CAD [264] for masses or even a decreased specificity without an improvement in detection rate or characterization of invasive cancers [79, 164]. Luckily, other studies show far more positive results suggesting an increase in detection rate of early stage malignant tumors [85] and an increase in sensitivity from 64% to 95% compared to independent double reading [249]. A potential limitation of these studies is that 'CAD' is often seen as a single entity and referred to as such. There is a large variety in the performance of different CAD systems and essentially any algorithm that gives some output for a medical image can be seen as CAD. The actual algorithm used in these studies is the most important factor, but typically not treated as such. In spite of these severe limitations, these studies are often cited and negatively affect the image of the general concept of CAD.

1.4.1 Applications of mammography CAD

Currently, CAD for mammography is used or proposed in roughly three different settings:

1. **CAD prompts** This is the setting in which CAD is originally proposed and the most common setting to this day. Readers of images are presented with CAD markers at suspicious locations for masses, calcifications or both. These systems are aimed at preventing false negatives due to oversight. Several variants are possible. Some systems show only markers, some show bounding boxes and some segmentations of potential malignancies. The way in which the information is presented may again influence the joint performance of the reader and the system.

2. **Interactive decision support** In this setting, instead of providing the reader with all findings above a certain threshold, images can be queried. These systems aim to support decision making, rather than prevent oversight errors [140, 226, 127].

3. **Independent second reader** CAD for mammography has also been proposed as an independent second reader or third reader [126]. Possible applications are to (1) select a small set of mammograms that received a high likelihood of malignancy by the CAD system but were not referred by the regular reading process and present these to an additional reader, or (2) use a system with high sensitivity as a pre-screening or filter and remove all mammograms that are obviously normal.

[7] http://www.cancer.org/healthy/findcancerearly/womenshealth/non-cancerousbreastconditions/non-cancerous-breast-conditions-fibroadenomas

In the development of these systems, different aspects need to be taken into account. For instance, support systems may need to be trained to be complementary and independent systems simply to outperform experienced readers, or operate close to 100% sensitivity while maintaining the best possible specificity.

1.5 Outline of this thesis

The main objective of this thesis is to introduce improved and novel algorithms for the detection of breast cancer, in particular soft-tissue lesions, in mammography. Given the success of deep learning in natural image analysis, all these methods are based on deep neural networks to some extent. The ultimate aim is a system that outperforms expert humans readers of mammograms and can operate independently.

The second chapter contains an introduction to deep learning in medical image analysis and relevant work on deep learning in mammography and other breast imaging applications that was incorporated into a larger survey we worked on. The third chapter presents a comparison between the CAD system that was under development in our lab and a plain deep convolutional neural network (CNN), a type of deep neural network particularly suitable for image analysis. Chapter 4 takes a brief sidestep from the deep learning track and presents a feature based system that was developed before deep learning became popular. The system discriminates cysts from solid lesions in a diagnostic setting. A deep CNN approach for the same problem is subsequently presented in chapter 5.

Chapter 6 presents methods that go beyond the standard patch-based approach of deep CNNs and add symmetrical and temporal information. In chapter 7, a method is presented that integrates all these sources of information along with findings output by a calcification detector. A summary of the entire thesis is provided in chapter 7. Every chapter contains a discussion about the method presented and suggestions for future work on the particular problem. A more general discussion on the implications of artificial intelligence in medicine and suggestions for future work is provided in chapter 8.

Chapter 2

Deep neural networks for medical image analysis

Relevant sections from
A Survey on Deep Learning in Medical Image Analysis - *Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, Clara I. Sánchez* - Medical Image Analysis, 2017.

Abstract

Deep learning algorithms, in particular convolutional networks, have rapidly become a methodology of choice for analyzing medical images. This paper reviews the major deep learning concepts pertinent to medical image analysis and summarizes over 300 contributions to the field, most of which appeared in the last year. We survey the use of deep learning for image classification, object detection, segmentation, registration, and other tasks and provide concise overviews of studies per application area.

2.1 Introduction

As soon as it was possible to scan and load medical images into a computer, researchers have built systems for automated analysis. Initially, from the 1970s to the 1990s, medical image analysis was done with sequential application of low-level pixel processing (edge and line detector filters, region growing) and mathematical modeling (fitting lines, circles and ellipses) to construct compound rule-based systems that solved particular tasks. There is an analogy with expert systems with many if-then-else statements that were popular in artificial intelligence in the same period. These expert systems have been described as GOFAI (good old-fashioned artificial intelligence) [111] and were often brittle; similar to rule-based image processing systems.

At the end of the 1990s, supervised techniques, where training data is used to develop a system, were becoming increasingly popular in medical image analysis. Examples include active shape models (for segmentation), atlas methods (where the atlases that are fit to new data form the training data), and the concept of feature extraction and use of statistical classifiers (for computer-aided detection and diagnosis). This pattern recognition or machine learning approach is still very popular and forms the basis of many successful commercially available medical image analysis systems. Thus, we have seen a shift from systems that are completely designed by humans to systems that are trained by computers using example data from which feature vectors are extracted. Computer algorithms determine the optimal decision boundary in the high-dimensional feature space. A crucial step in the design of such systems is the extraction of discriminant features from the images. This process is still done by human researchers and, as such, one speaks of systems with *handcrafted* features.

A logical next step is to let computers learn the features that optimally represent the data for the problem at hand. This concept lies at the basis of many deep learning algorithms: models (networks) composed of many layers that transform input data (e.g. images) to outputs (e.g. disease present/absent) while learning increasingly higher level features. The most successful type of models for image analysis to date are convolutional neural networks (CNNs). CNNs contain many layers that transform their input with convolution filters of a small extent. Work on CNNs has been done since the late seventies [90] and they were already applied to medical image analysis in 1995 by [176]. They saw their first successful real-world application in LeNet [162] for hand-written digit recognition. Despite these initial successes, the use of CNNs did not gather momentum until various new techniques were developed for efficiently training deep networks, and advances were made in core computing systems. The watershed was the contribution of [156] to the ImageNet challenge in December 2012. The proposed CNN, called AlexNet, won that competition by a large margin. In subsequent years, further progress has been made using related but deeper architectures [220]. In computer vision, deep convolutional networks have now become the technique of choice.

The medical image analysis community has taken notice of these pivotal developments. However, the transition from systems that use handcrafted features to systems that learn features from the data has been gradual. Before the breakthrough of AlexNet, many different techniques to learn features were popular. [10] provide a thorough review of these techniques. They include principal component analysis, clustering of image patches, dictionary approaches, and many more. [10] introduce CNNs that are trained end-to-end only at the end of their review in a section entitled *Global training of deep models*. In this survey, we focus particularly on such deep models, and do not include the more traditional feature learning approaches that have been applied to medical images. For a broader review on the application of deep learning in health informatics we refer to [214], where medical image analysis is briefly touched upon.

Applications of deep learning to medical image analysis first started to appear at workshops and conferences, and then in journals. The number of papers grew rapidly in 2015 and 2016. This is illustrated in Figure 2.1. The topic is now dominant at major conferences and a first special issue appeared of IEEE Transaction on Medical Imaging in May 2016 [106].

2.1.1 Learning algorithms

Machine learning methods are generally divided into *supervised* and *unsupervised learning* algorithms, although there are many nuances. In supervised learning, a model is presented with a dataset $\mathcal{D} = \{\mathbf{x}, y\}_{n=1}^{N}$ of input features $\mathbf{x}$ and label pairs y. This y can take several forms, depending on the learning task; in a classification setting y is generally a scalar representing a class label, whereas it can be a vector of continuous variables in the case if regression. When one tries to learn a segmentation model y can even be a multi-dimensional label image. Supervised training typically amounts

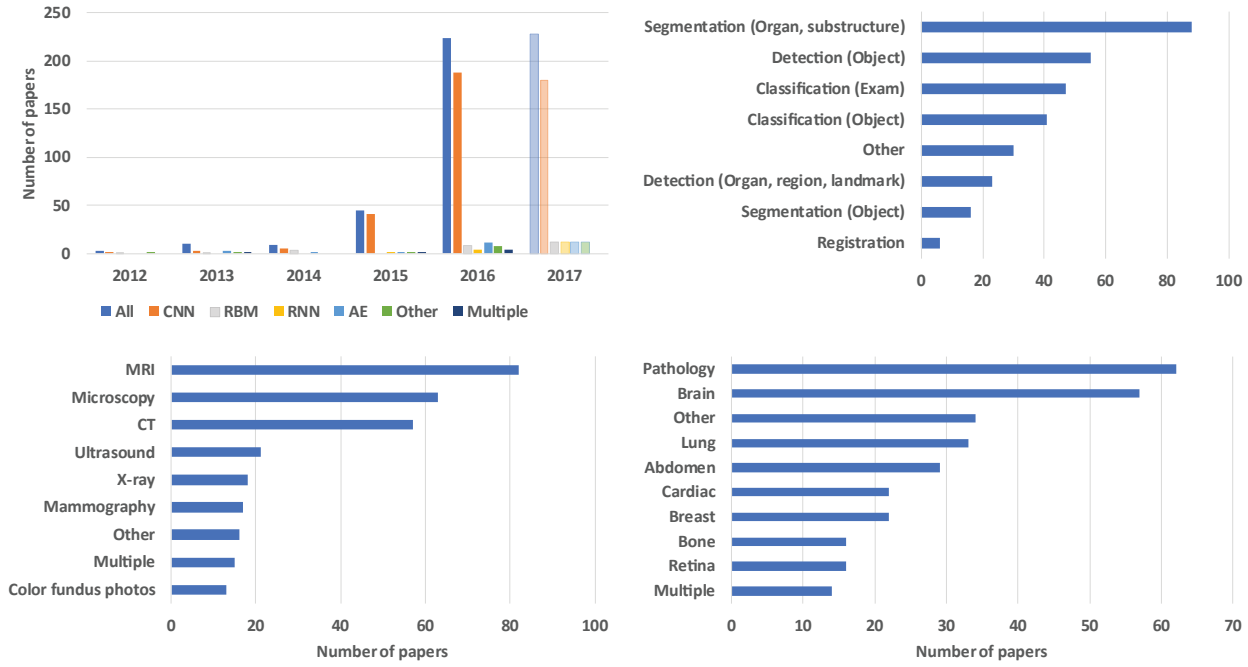


Figure 2.1: Breakdown of the papers included in this survey in year of publication, task addressed, imaging modality, and application area. The number of papers for 2017 has been extrapolated from the papers published in January.

to finding model parameters Θ that best predict the data based on a loss function $L(y, \hat{y})$. Here $\hat{y}$ denotes the output of the model obtained by feeding a data point $\mathbf{x}$ to the function $f(\mathbf{x}; \Theta)$ that represents the model.

Unsupervised learning algorithms process data without labels and are trained to find patterns, such as latent subspaces. Examples of traditional unsupervised learning algorithms are principal component analysis and clustering methods. Unsupervised training can be performed under many different loss functions. One example is reconstruction loss $L(\mathbf{x}, \hat{\mathbf{x}})$ where the model has to learn to reconstruct its input, often through a lower-dimensional or noisy representation.

2.1.2 Neural Networks

Neural networks are a type of learning algorithm which forms the basis of most deep learning methods. A neural network is comprised of neurons or units with some activation a and parameters $\Theta = \{\mathcal{W}, \mathcal{B}\}$, where $\mathcal{W}$ is a set of weights and $\mathcal{B}$ a set of biases. The activation represents a linear combination of the input $\mathbf{x}$ to the neuron and the parameters, followed by an element-wise non-linearity $\sigma(\cdot)$, referred to as a transfer function:

$$a = \sigma(\mathbf{w}^T \mathbf{x} + b) \tag{2.1}$$

Typical transfer functions for traditional neural networks are the sigmoid and hyperbolic tangent function. The multi-layered perceptrons (MLP), the most well-known of the traditional neural networks, have several layers of these transformations:

$$f(\mathbf{x}; \Theta) = \sigma(\mathbf{W}^L \sigma(\mathbf{W}^{L-1} \dots \sigma(\mathbf{W}^0 \mathbf{x} + b^0) + b^{L-1}) + b^L) \tag{2.2}$$

Here, $\mathbf{W^n}$ is a matrix comprising of rows $\mathbf{w}_k$, associated with activation k in the output. The symbol n indicates the number of the current layer, where L is the final layer. Layers in between the input and output are often referred to as 'hidden' layers. When a neural network contains multiple hidden layers it is typically considered a 'deep' neural network, hence the term 'deep learning'.

Often, the activations of the final layer of the network are mapped to a distribution over classes $P(y|\mathbf{x}; \Theta)$ through a *softmax* function:

$$P(y|\mathbf{x}; \Theta) = \text{softmax}(\mathbf{x}; \Theta) = \frac{e^{\left(\mathbf{w}_i^L\right)^T \mathbf{x} + b_i^L}}{\sum_{k=1}^{K} e^{\left(\mathbf{w}_k^L\right)^T \mathbf{x} + b_k^L}} \tag{2.3}$$

where $\mathbf{w}_i^L$ indicates the weight vector leading to the output node associated with class i.

Maximum likelihood with stochastic gradient descent is currently the most popular method to fit parameters Θ to a dataset $\mathcal{D}$. In stochastic gradient descent a small subset of the data, a mini-batch, is used for each gradient update instead of the full data set. Optimizing maximum likelihood in practice amounts to minimizing the negative log-likelihood:

$$\arg\min_{\Theta} -\sum_{n=1}^{N} \log\left[P(y_n|\mathbf{x}_n;\Theta)\right] \tag{2.4}$$

This results in the binary cross-entropy loss for two-class problems and the categorical cross-entropy for multi-class tasks. A downside of this approach is that it typically does not optimize the quantity we are interested in directly, such as area under the receiver-operating characteristic (ROC) curve or common evaluation measures for segmentation, such as the Dice coefficient.

For a long time, deep neural networks (DNN) were considered hard to train efficiently. They only gained popularity in 2006 [11, 115, 119] when it was shown that training DNNs layer-by-layer in an unsupervised manner (pre-training), followed by supervised fine-tuning of the stacked network, could result in good performance. Two popular architectures trained in such a way are stacked auto-encoders (SAEs) and deep belief networks (DBNs). However, these techniques are rather complex and require a significant amount of engineering to generate satisfactory results.

Currently, the most popular models are trained end-to-end in a supervised fashion, greatly simplifying the training process. The most popular architectures are convolutional neural networks (CNNs) and recurrent neural networks (RNNs). CNNs are currently most widely used in (medical) image analysis, although RNNs are gaining popularity. The following sections will give a brief overview of each of these methods, starting with the most popular ones, and discusses their differences and potential challenges when applied to medical problems.

2.1.3 Convolutional Neural Networks (CNNs)

There are two key differences between MLPs and CNNs. First, in CNNs weights in the network are shared in such a way that the network performs convolution operations on images. This way, the model does not need to learn separate detectors for the same object occurring at different positions in an image, making the network equivariant with respect to translations of the input. It also drastically reduces the amount of parameters (i.e. the number of weights no longer depends on the size of the input image) that need to be learned.

At each layer, the input image is convolved with a set of K kernels $\mathcal{W} = \{\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_K\}$ and added biases $\mathcal{B} = \{b_1, \ldots, b_K\}$, each generating a new feature map $\mathbf{X}_k$. These features are subjected to an element-wise non-linear transform $\sigma(\cdot)$ and the same process is repeated for every convolutional layer l:

$$\mathbf{X}_k^l = \sigma\big(\mathbf{W}_k^{l-1} * \mathbf{X}^{l-1} + b_k^{l-1}\big). \tag{2.5}$$

The second key difference between CNNs and MLPs, is the typical incorporation of pooling layers in CNNs, where pixel values of neighborhoods are aggregated using a permutation invariant function, typically the max or mean operation. This can induce a certain amount of translation invariance and increase the receptive field of subsequent convolutional layers. At the end of the convolutional stream of the network, fully-connected layers (i.e. regular neural network layers) are usually added, where weights are no longer shared. Similar to MLPs, a distribution over classes is generated by feeding the activations in the final layer through a softmax function and the network is trained using maximum likelihood.

2.1.4 Deep CNN Architectures

Given the prevalence of CNNs in medical image analysis, we elaborate on the most common architectures and architectural differences among the widely used models.

General classification architectures

LeNet [162] and AlexNet [156], introduced over a decade later, were 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 three 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. These deeper architectures generally have a lower memory footprint during inference, which enable their deployment on mobile computing devices such as smartphones. [246] were among the first to explore much deeper networks, and employed small, fixed size kernels in each layer. A 19-layer model often referred to as VGG19 or OxfordNet won the ImageNet challenge of 2014.

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. [260] introduced a 22-layer network named *GoogLeNet*, also referred to as Inception, which made use of so-called inception blocks [167], a module that replaces the mapping defined in Eq. (2.5) 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. The *ResNet* architecture [112] won the ImageNet challenge in 2015 and consisted of so-called ResNet-blocks. Rather than learning a function, the residual block only learns the residual and is thereby pre-conditioned towards learning mappings in each layer that are close to the identity function. This way, even deeper models can be trained effectively.

Since 2014, the performance on the ImageNet benchmark has saturated and it is difficult to assess whether the small increases in performance can really be attributed to 'better' and more sophisticated architectures. The advantage of the lower memory footprint these models provide is typically not as important for medical applications. Consequently, AlexNet or other simple models such as VGG are still popular for medical data, though recent landmark studies all use a version of GoogleNet called Inception v3 [108, 74, 175]. Whether this is due to a superior architecture or simply because the model is a default choice in popular software packages is again difficult to assess.

Multi-stream architectures

The default CNN architecture can easily accommodate multiple sources of information or representations of the input, in the form of channels presented to the input layer. This idea can be taken further and channels can be merged at any point in the network. Under the intuition that different tasks require different ways of fusion, *multi-stream* architectures are being explored. These models, also referred to as dual pathway architectures [136], have two main applications at the time of writing: (1) multi-scale image analysis and (2) 2.5D classification; both relevant for medical image processing tasks.

For the detection of abnormalities, context is often an important cue. The most straightforward way to increase context is to feed larger patches to the network, but this can significantly increase the amount of parameters and memory requirements of a network. Consequently, architectures have been investigated where context is added in a down-scaled representation in addition to high-resolution local information. To the best of our knowledge, the multi-stream multi-scale architecture was first explored by [77], who used it for segmentation in natural images. Several medical applications have also successfully used this concept [136, 186, 251, 289].

As so much methodology is still developed on natural images, one of the challenges of applying deep learning techniques to the medical domain often lies in adapting existing architectures to, for instance, different input formats such as three-dimensional data. In early applications of CNNs to such volumetric data, full 3D convolutions and the resulting large amount of parameters were circumvented by dividing the Volume of Interest (VOI) into slices which are fed as different streams to a network. [207] were the first to use this approach for knee cartilage segmentation. Similarly, the network can be fed with multiple angled patches from the 3D-space in a multi-stream fashion, which has been applied by various authors in the context of medical imaging [218, 233]. These approaches are also referred to as 2.5D classification.

Segmentation Architectures

Segmentation is a common task in both natural and medical image analysis and to tackle this, CNNs can simply be used to classify each pixel in the image individually, by presenting it with patches extracted around the particular pixel. A drawback of this naive 'sliding-window' approach is that input patches from neighboring pixels have huge overlap and the same convolutions are computed many times. Fortunately, the convolution and dot product are both 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 take input images larger than it was trained on and produce a likelihood map, rather than an output for a single pixel. The resulting 'fully convolutional network' (fCNN) 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' [178] is one of several methods proposed to prevent this decrease in resolution. The fCNN is applied to shifted

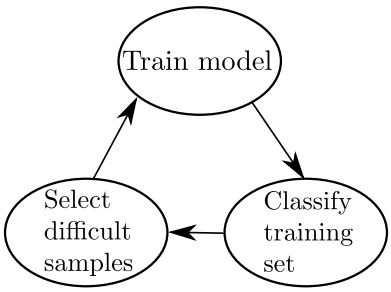


Figure 2.2: Illustration of the cascaded approach. The work in this thesis generally has two iterations, a candidate detection phase and a classification phase.

versions of the input image. By stitching the result together, one obtains a full resolution version of the final output, minus the pixels lost due to the 'valid' convolutions.

[217] took the idea of the fCNN one step further and proposed the U-net architecture, comprising a 'regular' fCNN followed by an upsampling part where 'up'-convolutions are used to increase the image size, coined contractive and expansive paths. Although this is not the first paper to introduce learned upsampling paths in convolutional neural networks (e.g. [178]), the authors combined it with so called skip-connections to directly connect opposing contracting and expanding convolutional layers. A similar approach was used by [44] for 3D data. [184] 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.

Deep structured architectures

A competing alternative to the U-Net like architecture is the addition of a layer on top of the posteriors output by a CNN, that models interaction between pixels or objects in an image. Markov Random Fields (MRF) are often used to finetune segmentations by taking correlations between neighboring structures into account. Interest in MRFs and in particular Conditional Random Fields (CRF) was recently reinvigorated when several groups proposed methods to jointly train deep CNNs and CRFs [40, 230, 299, 166]. Rather than fitting a posterior distribution $P(y|\mathbf{x};\Theta)$ over a single pixel or region of interest, they learn a joint distribution over a set of random variables $y_1, y_2, \ldots, y_K$:

$$P(y_1, y_2, \ldots, y_K|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K;\Theta) \tag{2.6}$$

This type of architecture currently claims best segmentation performance for the PASCAL VOC 2012 and NYUDv2 segmentation benchmarks in natural images [166]. Several medical applications are now incorporating a CRF to refine CNN segmentations [93, 136, 211].

Cascade architectures

Class imbalance is often a far more important hurdle in medical data than it is in natural images. Most diseases are rare and if they are present, they can be in the form of small abnormalities in large images. Discriminative models, including DNNs fare poorly, unless adjusted for class imbalance. Additionally, many 'normal' samples may convey negligible class discriminative information and therefore including them during training only slows down the learning process. *Cascaded architectures* are employed to tackle these issues. Many traditional systems before the deep learning revolution, in particular in CAD, used two stage pipelines. A first stage candidate detector is used to weed out an initial set of irrelevant locations, features were then computed from the relevant sites and subjected to a second classification stage.

This concept can be taken a step further by simply adding an iterative layer around the learning algorithm (and perform N stages instead of 2) that classifies the training set at each iteration. The set is resampled or weighted and the model is retrained. This strategy is referred to as *selective sampling* or *hard-negative mining* in the object detection community and has seen successful application to medical data [47, 279, 190]. The idea is related to boosting algorithms such as AdaBoost [86], in which case multiple models are averaged, each one trained on a weighted portion of the dataset that was found difficult by the ensemble of the previous iteration. An illustration of the cascade architecture is provided in figure 2.2. A challenge in using these cascades is that the system sometimes forgets what it learned in previous iterations.

2.1.5 Recurrent Neural Networks (RNNs)

Traditionally, RNNs were developed for discrete sequence analysis. They can be seen as a generalization of MLPs because both the input and output can be of varying length, making them suitable for tasks such as machine translation where a sentence of the source and target language are the input and output. In a classification setting, the model learns a distribution over classes $P(y|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T; \Theta)$ given a sequence $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$, rather than a single input vector $\mathbf{x}$.

The plain RNN maintains a latent or hidden state $\mathbf{h}$ at time t that is the output of a non-linear mapping from its input $\mathbf{x}_t$ and the previous state $\mathbf{h}_{t-1}$:

$$\mathbf{h}_t = \sigma(\mathbf{W}\mathbf{x}_t + \mathbf{R}\mathbf{h}_{t-1} + \mathbf{b}), \tag{2.7}$$

where weight matrices $\mathbf{W}$ and $\mathbf{R}$ are shared over time. For classification, one or more fully-connected layers are typically added followed by a softmax to map the sequence to a posterior over the classes.

$$P(y|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T; \Theta) = \text{softmax}(\mathbf{h}_T; \mathbf{W}_{out}, \mathbf{b}_{out}). \tag{2.8}$$

Since the gradient needs to be backpropagated from the output through time, RNNs are inherently deep (in time) and consequently suffer from the same problems with training as regular deep neural networks [14]. To this end, several specialized memory units have been developed, the earliest and most popular being the Long Short Term Memory (LSTM) cell [120]. The Gated Recurrent Unit [43] is a recent simplification of the LSTM and is also commonly used.

Although initially proposed for one-dimensional input, RNNs are increasingly applied to images. In natural images 'pixelRNNs' are used as autoregressive models, generative models that can eventually produce new images similar to samples in the training set. For medical applications, they have been used for segmentation problems, with promising results [253] in the MRBrainS challenge.

2.1.6 Pretraining and transfer learning

Early on in the development of deep models, the computer vision community discovered that the filters learned by CNNs are rather generic and can be transfered to other tasks, by only retraining the last layer of the network and obtain excellent results [293, 198, 215, 45]. Alternatively, the full network and all filters can be finetuned for the target task. This transferability of knowledge has a major advantage if tasks are related and data in the target task is scarce.

Filters learned from natural images transfer surprisingly well to medical data. The Overfeat model [215] was one of the earliest CNN applications to medical data and was shown to outperform previous work [278, 46, 30]. Pipelines based on pretrained CNNs from ImageNet claim top positions on leaderboards of challenges (reference to pathology challenge). Potential challenges in the application of nets trained on natural images are finding the optimal point at which to stop fine-tuning (elaborate), the best scaling of input patches and handling grey scale data in a model trained on three channel color images. Currently, medical applications employing pretrained models from ImageNet are ubiquitous [173, 158, 93, 143, 94, 188, 235, 205, 41, 193, 38]. Shin et al. [236] investigated the effect of pretraining and report that it consistently performs on par or outperforms models trained from scratch. Tashbakhsh et al. [263] come to a similar conclusion and study the depth to which networks need to be retrained for various medical tasks.

2.1.7 Unsupervised models

Auto-encoders (AEs) and Stacked Auto-encoders (SAEs)

AEs are simple networks that are trained to reconstruct the input $\mathbf{x}$ on the output layer $\mathbf{x}'$ through one hidden layer $\mathbf{h}$. They are governed by a weight matrix $\mathbf{W}_{x,h}$ and bias $b_{x,h}$ from input to hidden state and $\mathbf{W}_{h,x'}$ with corresponding bias $b_{h,x'}$ from the hidden layer to the reconstruction. A non-linear function is used to compute the hidden activation:

$$\mathbf{h} = \sigma(\mathbf{W}_{x,h}\mathbf{x} + \mathbf{b}_{x,h}). \tag{2.9}$$

Additionally, the dimension of the hidden layer $|\mathbf{h}|$ is taken to be smaller than $|\mathbf{x}|$. This way, the data is projected onto a lower dimensional subspace representing a dominant latent structure in the input. Regularization or sparsity constraints can be employed to enhance the discovery process. If the hidden layer had the same size as the input and no further non-linearities were added, the model would simply learn the identity function.

The denoising auto-encoder [284] is another solution to prevent the model from learning a trivial solution. Here the model is trained to reconstruct the input from a noise corrupted version (typically salt-and-pepper-noise). SAEs (or deep

AEs) are formed by placing auto-encoder layers on top of each other. In medical applications surveyed in this work, auto-encoder layers were often trained individually ('greedily') after which the full network was fine-tuned using supervised training to make a prediction.

Restricted Boltzmann Machines (RBMs) and Deep Belief Networks (DBNs)

RBMs [116] are a type of Markov Random Field (MRF), constituting an input layer or visible layer $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ and a hidden layer $\mathbf{h} = (h_1, h_2, \ldots, h_M)$ that carries the latent feature representation. The connections between the nodes are bi-directional, so given an input vector $\mathbf{x}$ one can obtain the latent feature representation $\mathbf{h}$ and also vice versa. As such, the RBM is a generative model, and we can sample from it and generate new data points. In analogy to physical systems, an energy function is defined for a particular state $(\mathbf{x}, \mathbf{h})$ of input and hidden units:

$$E(\mathbf{x}, \mathbf{h}) = \mathbf{h}^T \mathbf{W} \mathbf{x} - \mathbf{c}^T \mathbf{x} - \mathbf{b}^T \mathbf{h}, \tag{2.10}$$

with $\mathbf{c}$ and $\mathbf{b}$ bias terms. The probability of the 'state' of the system is defined by passing the energy to an exponential and normalizing:

$$p(\mathbf{x}, \mathbf{h}) = \frac{1}{Z} \exp\{-E(\mathbf{x}, \mathbf{h})\}. \tag{2.11}$$

Computing the partition function Z is generally intractable. However, conditional inference in the form of computing $\mathbf{h}$ conditioned on $\mathbf{x}$ or vice versa is tractable and results in a simple formula:

$$P(h_j|\mathbf{x}) = \frac{1}{1 + \exp\{-b_j - \mathbf{W}_j \mathbf{x}\}}. \tag{2.12}$$

Since the network is symmetric, a similar expression holds for $P(x_i|\mathbf{h})$.

DBNs [11, 119] are essentially SAEs where the AE layers are replaced by RBMs. Training of the individual layers is, again, done in an unsupervised manner. Final fine-tuning is performed by adding a linear classifier to the top layer of the DBN and performing a supervised optimization.

Variational Auto-Encoders and Generative Adverserial Networks

Recently, two novel unsupervised architectures were introduced: the variational auto-encoder (VAE) [145] and the generative adversarial network (GAN) [102]. There are no peer-reviewed papers applying these methods to medical images yet, but applications in natural images are promising. We will elaborate on their potential in the discussion.

2.1.8 Hardware and Software

One of the main contributors to the steep rise of deep learning papers has been the widespread availability of GPU and GPU-computing libraries (CUDA, OpenCL). GPUs are highly parallel computing engines, which have an order of magnitude more execution threads than central processing units (CPUs). With current hardware, deep learning on GPUs is typically 10 to 30 times faster than on CPUs.

Next to hardware, the other driving force behind the popularity of deep learning methods is the wide availability of open-source software packages. These libraries provide efficient GPU implementations of important operations in neural networks, such as convolutions; allowing the user to implement ideas at a high level rather than worrying about efficient implementations. At the time of writing, the most popular packages were (in alphabetical order):

- **Caffe** [134]. Provides C++ and Python interfaces, developed by graduate students at UC Berkeley.
- **Tensorflow** [1]. Provides C++ and Python and interfaces, developed by Google and is used by Google research.
- **Theano** [9]. Provides a Python interface, developed by MILA lab in Montreal.
- **Torch** [52]. Provides a Lua interface and is used by, among others, Facebook AI research.

There are third-party packages written on top of one or more of these frameworks, such as Lasagne (`https://github.com/Lasagne/Lasagne`) or Keras (`https://keras.io/`). It goes beyond the scope of this paper to discuss all these packages in detail.

2.2 Deep Learning in Breast Image Analysis

One of the earliest DNN applications from [223] was on breast imaging. Recently, interest has returned which resulted in significant advances over the state of the art, achieving the performance of human readers on ROIs [153]. Since most breast imaging techniques are two dimensional, methods successful in natural images can easily be transferred. With one exception, the only task addressed is the detection of breast cancer; this consisted of three subtasks: (1) detection and classification of mass-like lesions, (2) detection and classification of micro-calcifications, and (3) breast cancer risk scoring of images. Mammography is by far the most common modality and has consequently enjoyed the most attention. Work on tomosynthesis, US, and shear wave elastography is still scarce, and we have only one paper that analyzed breast MRI with deep learning; these other modalities will likely receive more attention in the next few years. Table 2.1 summarizes the literature and main messages.

Since many countries have screening initiatives for breast cancer, there should be massive amounts of data available, especially for mammography, and therefore enough opportunities for deep models to flourish. Unfortunately, large public digital databases are still unavailable and consequently older scanned screen-film data sets such as the MIAS or DDSM are still in use. Challenges such as the recently launched DREAM challenge have not yet had the desired success.

As a result, many papers used small data sets resulting in mixed performance. Several projects have addressed this issue by exploring semi-supervised learning [256], weakly supervised learning [129], and transfer learning [154, 225]). Another method combines deep models with handcrafted features [59], which have been shown to be complementary still, even for very big data sets [153]. State of the art techniques for mass-like lesion detection and classification tend to follow a two-stage pipeline with a candidate detector; this design reduces the image to a set of potentially malignant lesions, which are fed to a deep CNN [82, 153]. Alternatives use a region proposal network (R-CNN) that bypasses the cascaded approach [3, 147].

Table 2.1: Overview of papers using deep learning techniques for breast image analysis. MG = mammography; TS = tomosynthesis; US = ultrasound; ADN = Adaptive Deconvolution Network.

Reference	Modality	Method	Application; remarks
[223]	MG	CNN	First application of a CNN to mammography
[132]	MG, US	ADN	Four layer ADN, an early form of CNN for mass classification
[81]	MG	CNN	Pre-trained network extracted features classified with an SVM for breast density estimation
[3]	MG	CNN	Use a modified region proposal CNN (R-CNN) for the localization and classification of masses
[5]	MG	CNN	Lesion classification, combination with hand-crafted features gave the best performance
[54]	MRI	CNN	Breast and fibroglandular tissue segmentation
[67]	MG	CNN	Tissue classification using regular CNNs
[59]	MG	CNN	Combination of different CNNs combined with hand-crafted features
[82]	TS	CNN	Improved state-of-the art for mass detection in tomosynthesis
[129]	MG	CNN	Weakly supervised CNN for localization of masses
[128]	MG	CNN	Pre-trained CNN on natural image patches applied to mass classification
[135]	MG	SAE	Unsupervised CNN feature learning with SAE for breast density classification
[147]	MG	CNN	R-CNN combined with multi-class loss trained on semantic descriptions of potential masses
[153]	MG	CNN	Improved the state-of-the art for mass detection and show human performance on a patch level
[209]	MG	CNN	CNN for direct classification of future risk of developing cancer based on negative mammograms
[224]	TS	CNN	Microcalcification detection
[225]	TS	CNN	Pre-trained CNN on mammographic masses transfered to tomosynthesis
[256]	MG	CNN	Semi-supervised CNN for classification of masses
[297]	US	RBM	Classification benign vs. malignant with shear wave elastography
[154]	MG	CNN	Pre-trained CNN on mass/normal patches to discriminate malignant masses from (benign) cysts
[286]	MG	CNN	Detection of cardiovascular disease based on vessel calcification

Chapter 3

Detection of masses in mammograms using a deep convolutional neural network

Appeared in:
Large Scale Deep Learning for Computer Aided Detection of Mammographic Lesions - *Thijs Kooi, Geert Litjens, Bram van Ginneken, Albert Gubern-Mérida, Clara I. Sánchez, Ritse Mann, Ard den Heeten and Nico Karssemeijer* - Medical Image Analysis, 2017

Abstract

Recent advances in machine learning yielded new techniques to train deep neural networks, which resulted in highly successful applications in many pattern recognition tasks such as object detection and speech recognition. In this chapter we provide a head-to-head comparison between a state-of-the art in mammography CAD system, relying on a manually designed feature set and a Convolutional Neural Network (CNN), aiming for a system that can ultimately read mammograms independently. Both systems are trained on a large data set of around 45000 images and results show the CNN outperforms the traditional CAD system at low sensitivity and performs comparable at high sensitivity. We subsequently investigate to what extent features such as location and patient information and commonly used manual features can still complement the network and see improvements at high specificity over the CNN especially with location and context features, which contain information not available to the CNN. Additionally, a reader study was performed, where the network was compared to certified screening radiologists on a patch level and we found no significant difference between the network and the readers.

3.1 Introduction

Until recently, the effectiveness of CAD systems and many other pattern recognition applications depended on meticulously handcrafted features, topped off with a learning algorithm to map it to a decision variable. Radiologists are often consulted in the process of feature design and features such as the contrast of the lesion, spiculation patterns and the sharpness of the border are used, in the case of mammography. These feature transformations provide a platform to instill task-specific, a-priori knowledge, but cause a large bias towards how we humans think the task is performed. Since the inception of Artificial Intelligence (AI) as a scientific discipline, research has seen a shift from rule-based, problem specific solutions to increasingly generic, problem agnostic methods based on learning, of which *deep learning* [12, 10, 229, 163] is its most recent manifestation. Directly distilling information from training samples, rather than the domain expert, deep learning allows us to optimally exploit the ever increasing amounts of data and reduce human bias. For many pattern recognition tasks, this has proven to be successful to such an extent that systems are now reaching human or even superhuman performance [48, 185, 113].

To the best of our knowledge, Sahiner et al. [223] were the first to attempt a CNN setup for mammography. Instead of raw images, texture maps were fed to a simple network with two hidden layers, producing two and three feature images respectively. The method gave acceptable, but not spectacular results. Many things have changed since this publication, however, not only with regard to statistical learning, but also in the context of acquisition techniques. Screen Film Mammography (SFM) has made way for Digital Mammography (DM), enabling higher quality, raw images in which pixel values have a well-defined physical meaning and easier spread of large amounts of training data. Given the advances in learning and data, we feel a revisit of CNNs for mammography is more than worthy of exploration.

In previous work in our group [126] we showed that a sophisticated CAD system taking into account not only local information, but also context, symmetry and the relation between the two views of the same breast can operate at the performance of a resident radiologist and of a certified radiologist at high specificity. In a different study [139] it was shown that when combining the judgment of up to twelve radiologists, reading performance improved, providing a lower bound on the maximum amount of information in the medium and suggesting ample room for improvement of the current system.

In this chapter, we provide a head-to-head comparison between a CNN and a CAD system relying on an exhaustive set of manually designed features and show the CNN outperforms a state-of-the-art mammography CAD system, trained on a large dataset of around 45000 images. We will focus on the detection of solid, malignant lesions including architectural distortions, treating benign abnormalities such as cysts or fibroadenomae as false positives. The goal of this chapter is *not* to give an optimally concise set of features, but to use a complete set where all descriptors commonly applied in mammography are represented and provide a fair comparison with the deep learning method. As mentioned by Szegedy et al. [260], success in the past two years in the context of object recognition can in part be attributed to judiciously combining CNNs with classical computational vision techniques. In this spirit, we employ a candidate detector to obtain a set of suspicious locations, which are subjected to further scrutiny, either by the classical system or the CNN. We subsequently investigate to what extent the CNN is still complementary to traditional descriptors by combining the learned representation with features such as location, contrast and patient information, part of which are not explicitly represented in the patch fed to the network. Lastly, a reader study is performed, where we compare the scores of the CNN to experienced radiologists on a patch level.

The rest of this chapter is organized as follows. In the next section, we will give details regarding the candidate detection system, shared by both methods. In section 6.2.1, the CNN will be introduced followed by a description of the reference system in section 3.3. In section 3.4, we will describe the experiments performed and present results, followed by a discussion in section 3.5 and conclusion in section 3.6.

3.2 Candidate Detection

Before gathering evidence, every pixel is a possible center of a lesion. This approach yields few positives and an overwhelming amount of predominantly obvious negatives. The actual difficult examples could be assumed to be outliers and generalized away, hindering training. Sliding window methods, previously popular in image analysis are recently losing ground in favor of candidate detection [122] such as selective search [273] to reduce the search space [100, 260]. We therefore follow a two-stage classification procedure where in the first stage, candidates are detected and subjected to

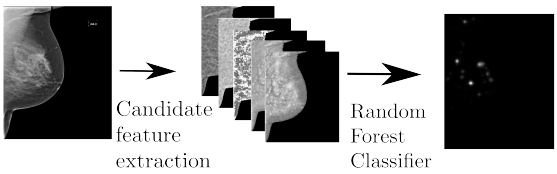
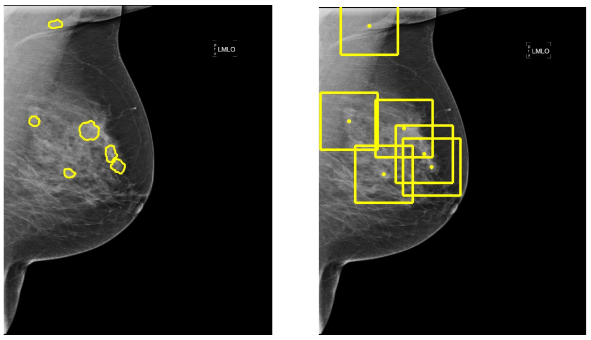


Figure 3.1: Illustration of the candidate detection pipeline. A candidate detector is trained using five pixel features and applied to all pixels in all images, generating a likelihood image. Local optima in the likelihood image are used as seed points for both the reference system and the CNN (see figure 3.2).



(a) Illustration of segmentations for the reference system.

(b) Illustration of extracted patches for the CNN.

Figure 3.2: Two systems are compared. A candidate detector (see figure 3.1) generates a set of candidate locations. A traditional CAD system (left) uses these locations as seed points for a segmentation algorithm. The segmentations are used to compute region based features. The second system based on a CNN (right) uses the same locations as the center of a region of interest.

further scrutiny in a second stage, similar to the pipeline described in [126]. Rather than class agnostic and potentially less accurate candidate detection methods, we use an algorithm designed for mammographic lesions [142]. It operates by extracting five features based on first and second order Gaussian kernels, two designed to spot the center of a focal mass and two looking for spiculation patterns, characteristic of malignant lesions. A final feature indicates the size of optimal response in scale-space.

To generate the pixel based training set, we extracted positive samples from a disk of constant size inside each annotated malignant lesion in the training set, to sample the same amount from every lesion size and prevent bias for larger areas. To obtain normal pixels for training, we randomly sampled 1 in 300 pixels from normal tissue in normal images, resulting in approximately 130 negative samples per normal image. The resulting samples were used to train a random forest [24] (RF) classifier. RFs can be parallelized easily and are therefore fast to train, are less susceptible to overfitting and easily adjustable for class-imbalance and therefore suitable for this task.

To obtain lesion candidates, the RF is applied to all pixel locations in each image, both in the train and test set, generating a likelihood image, where each pixel indicates the estimated suspiciousness. Non-maximum suppression was performed on this image and all optima in the likelihood image are treated as candidates and fed as input to both the reference feature system and the CNN. For the reference system, the local optima in the likelihood image are used as seed points for a segmentation algorithm. For the CNN, a patch centered around the location is extracted. An overview of the first stage pipeline is provided in figure 3.1. Figure 3.2 illustrates the generated candidates for both systems.

3.2.1 Data Augmentation

Although powerful, contemporary architectures are not fully invariant to geometric transformations, such as rotation and scale. Data augmentation is a technique often used in the context of deep learning and refers to the process of generating new samples from data we already have, hoping to ameliorate data scarcity and prevent overfitting. In object recognition tasks in natural images, simple horizontal flipping is usually only performed, but for tasks such as optical character recog-

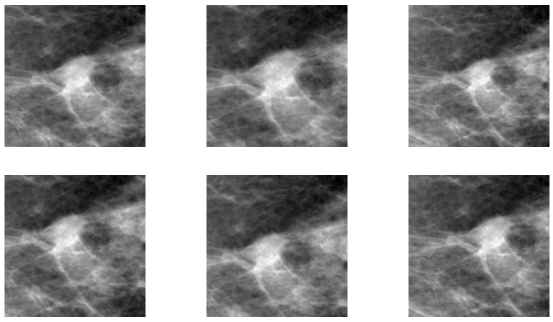


Figure 3.3: Examples of scaling and translation of the patches. The top left image is the original patch, the second and third image of the top row examples of the smallest and largest scaling employed. The bottom row indicates the extrema in the range of translation used.

nition it has been shown that elastic deformations can greatly improve performance [244]. The main sources of variation in mammography at a lesion level are rotation, scale, translation and the amount of occluding tissue.

We augmented all positive examples with scale and translation transformations. Full scale or translation invariance is not desired nor required since the candidate detector is expected to find a patch centered around the actual focal point of the lesion. The problem is not completely scale-invariant either: large lesions in a later stage of growth are not simply scaled-up versions of recently emerged abnormalities. The key is therefore to perform the right amount of translation and scaling in order to generate realistic lesion candidates. To this end, we translate each patch in the training set containing an annotated malignant lesion 16 times by adding values sampled uniformly from the interval $[-25, 25]$ (0.5 cm) to the lesion center and scale it 16 times by adding values from the interval $[-30, 30]$ (0.6 cm) to the top left and bottom right of the bounding box. After this, all patches including the normals were rotated using simple flipping actions, which can be computed on the fly to generate three more samples. This results in $(1 + 16 + 16)4 = 132$ patches per positive lesions and 4 per negative. Examples of the range of scaling and translation augmentation are given in figure 3.3.

3.3 Reference System

The large majority of CAD systems rely on some form of segmentation of the candidates on which region based features are computed. To this end, we employ the mass segmentation method proposed by Timp and Karssemeijer [271], which was shown to be superior to other methods (region growing [265] and active contour segmentation [160]) on their particular feature set. The image is transformed to a polar domain around the center of the candidate and dynamic programming is used to find an optimal contour, subject to the constraint that the path must start and end in the same column in order to generate a closed contour in the Cartesian domain. A cost function incorporating a deviation from the expected gray level, edge strength and size terms is used to find an optimal segmentation. One of the problems with this method and many knowledge driven segmentation methods for that matter, is that it is conditioned on a *false prior*: the size constraint is based on data from malignant lesions. When segmenting a candidate, we therefore implicitly assume that this is a malignant region, inadvertently driving the segmentation into a biased result. Many of the manual features described below rely on a precise segmentation but in the end, it is an intermediate problem. For a stand-alone application, we are interested to provide the patient with an accurate diagnosis, not the exact delineation. A huge advantage of CNNs is that no segmentation is required and patches are fed without any intermediate processing.

After segmentation, we extract a set of 74 features. These can broadly be categorized into *pixel level features*, used by the candidate detector, *contrast features*, capturing the relation between the attenuation coefficients inside and outside the region, *texture features* describing relations between pixels within the segmented region, *geometry features* summarizing shape and border information *location* features, indicating where the lesion is with respect to some landmarks in the breast, *context* features, capturing information about the rest of the breast and other candidates and *patient* features, conveying some of the subjects background information.

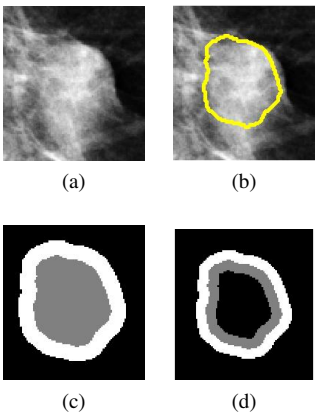


Figure 3.4: A lesion (a), its segmentation (b), areas used for computing contrast features (c) and areas used for computing margin contrast (d)

3.3.1 Candidate Detector Features

As a first set of descriptors, we re-use the five features employed by the candidate detector, which has been shown to be beneficial in previous work in our group. On top of this, we compute the mean of the four texture features within the segmented boundary and add the output of the candidate detector at the found optimum. This gives us a set of nine outputs we call candidate detector features.

3.3.2 Contrast Features

When talking to a radiologist, a feature that is often mentioned is how well a lesion is separated from the background. Contrast features are designed to capture this. To compute these, we apply a distance transform to the segmented region and compare the inside of the segmentation with a border around it. The distance d to the border of the segmentation is determined according to:

$$d = \rho\sqrt{A\pi} \tag{3.1}$$

with A the area of the segmented lesion. An illustration is provided in figure 3.4. An important nuisance in this setting is the tissue surrounding the lesion. In previous work, we have derived two model based features, designed to be invariant to this factor [150], which were also normalized for size of the lesion. The *sharpness* of the border of the lesion is also often mentioned by clinicians. To capture this, we add two features: the acutance [212] and margin contrast, the different between the inside and outside of the segmentation, using a small margin. Illustrations of contrast features are provided in figure 3.4. Other contrast features described in [266] were added to give a set of 12 features.

3.3.3 Texture Features

The presence of holes in the candidate lesion often decrease their suspiciousness, since tumours are solid, with possibly the exception of lobular carcinoma. To detect this, we added the two isodensity features proposed by Te Brake and Karssemeijer [266]. Linear structures within a lesion can indicate an unfortunate projection rather then cancer, for which we used four linear texture features as described by the same authors [266]. On top of this we added two features based on the second order gradient image of the segmented lesion. The image was convolved with second order Gaussian derivative filters and the optimal location in scale space was selected for each pixel. We subsequently took the first and second moment of the segmented lesion of the maximum magnitude, which is expected to be high for lesions with much line structure. Secondly, we computed gradient coocurence, by counting the number of times adjacent pixels have the same orientation. Ten less biophysical features in the form of Haralick features [110] at two different scales (*entropy*, *contrast*, *correlation*, *energy* and *homogeneity*) were added to give a set of 21 texture descriptors.

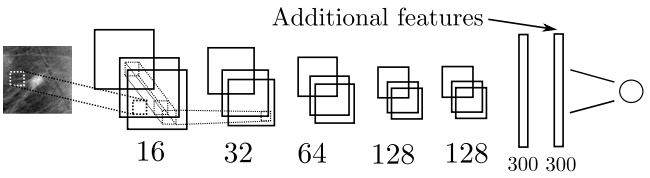


Figure 3.5: Illustration of the network architecture, The numbers indicate the amount of kernels used. We employ a scaled down version of the VGG model. To see the extent to which conventional features can still help, the network is trained fully supervised and the learned features are subsequently extracted from the final layer and concatenated with the manual features and retrained using a second classifier.

3.3.4 Geometrical Features

Regularity of the border of a lesion is often used to classify lesions. Again, expedient computation relies heavily on proper segmentations. Nevertheless, we have incorporated five simple topology descriptors as proposed by [204] in the system. These are *eccentricity*, *convexity*, *compactness*, *circular variance* and *elliptic variance*. In order to capture more of the 3D shape, we extended these descriptors to also work with 3 dimensions. The lesion was smoothed with a Gaussian kernel first and *3D eccentricity*: the ratio between the largest and smallest eigenvalue of the point cloud, *3D compactness*: the ratio of the surface area to the volume, *spherical deviance*, the average deviation of each point from a sphere and *elliptical deviance*: the average deviation of each point to an ellipse fitted to the point cloud were computed. Since convex hull algorithms in 3D suffer from relatively high computational complexity, this was not extended. Op top of this, we added a feature measuring reflective symmetry. The region is divided into radial and angular bins and average difference pixel intensity between opposing bins is summed and normalized by the size of the region. Lastly the area of the segmented region is added, giving us a set of 10 geometric features.

3.3.5 Location Features

Lesions are more likely to occur in certain parts of the breast than others and other structures such as lymph nodes are more common in the pectoralis than in other parts of the breast. To capture this, we use a simple coordinate system. The area of the breast and pectoral muscle are segmented using thresholding and a polynomial fit. We subsequently estimate the nipple location by taking the largest distance to the chest wall and a central landmark in the chest wall is taken as the row location of the center of gravity. From this, we extract: (1) the distance to the nipple (2) the same, but normalized for the size of the breast, (3) the distance to the chest wall and (4) the fraction of the lesion that lies in the pectoral muscle.

3.3.6 Context Features

To add more information about the surroundings of the lesion, we added three context features as described by Hupse and Karssemeijer [124]. The features again make use of the candidate detector and assume the posterior of pixels in the rest of the breast convey some information about the nature of the lesion in question. The first feature averages the output around the lesion, the second in a band at a fixed distance from the nipple and a third takes the whole segmented breast into account. On top of this, we added the posterior of the candidate detector, normalized by the sum of the top three and top five lesions in the breast, to give us five context features in total.

3.3.7 Patient Features

Lastly, we added the age of the patient, which is an important risk factor. From the age, we also estimate the screening round by subtracting 50 (the age at which screening starts in The Netherlands) and dividing by 2 (the step size of the screening). This gives us two features.

Note that the last three sets of features provide information outside of the patch fed to the CNN. Even if the network is able to exploit all information in the training set, these could still supply complementary information regarding the nature of the lesion.

Table 3.1: Overview of the data. Pos refers to the amount of malignant lesions and neg to the amount of normals.

	Cases		Exams		Images		Candidates	
	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
Train	296	6433	358	11780	634	39872	634	213450
Valid.	35	710	42	1247	85	4218	85	19460
Test	124	2064	124	5317	271	18182	271	180777

3.4 Experiments

3.4.1 Data

The mammograms used were collected from a large scale screening program in The Netherlands (*bevolkingsonderzoek midden-west*) and recorded using a Hologic Selenia digital mammography system. All tumours are biopsy proven malignancies and annotated by an experienced reader. Before presentation to a radiologist, the manufacturer applies some processing to optimize it for viewing by a human. To prevent information loss and bias, we used the raw images instead and only applied a log transform which results in pixel values being linearly related to the attenuation coefficient. Images were scaled from 70 micron to 200 for faster processing. Structure important for detecting lesions occur at larger scales and therefore this does not cause any loss of information.

An overview of the data is provided in table 7.1. With the term case, we refer to all screening images recorded from a single patient. Each case consists of several exams taken at typically a two year interval and each exam typically comprises four views, two of each breast, although these numbers vary: some patients skip a screening and for some exams only one view of each breast is recorded. For training and testing, we selected all regions found by the candidate detector. The train, validation and test set were all split on a patient level to prevent any bias. The train and validation set comprise 44090 mammographic views, from which we used 39872 for training and 4218 for validation. The test set consisted of 18182 images of 2064 patients with 271 malignant annotated lesions. A total of 30 views from 20 exams in the test set contained an interval cancer that was visible in the mammogram or were taken prior to a screen detected cancer, with the abnormality already visible.

Before patch extraction in the CNN system, we segmented all lesions in the training set in order to get the largest possible lesion and choose the patch size with an extra margin resulting in patches of size 250×250 (5×5 cm). The pixel values in the patches were scaled using simple min-max scaling, with values calculated over the whole training set. We experimented with scaling the patches locally, but this seemed to perform slightly though not significantly worse on the validation set. All interpolation processes were done with bilinear interpolation. Since some candidates occur at the border of the imaged breast, we pad the image with zeros. Negative examples were only taken from normal images. Annotated benign samples such as cysts and fibroadenomae were removed from the training set. However, not all benign lesions in our data are annotated and therefore some may have ended in the train or validation set as negatives. After augmentation, the train set consisted of 334752 positive patches and 853800 negatives. When combining the train and validation set, this amounts to 379632 positive and 931640 negative patches.

3.4.2 Training and Classification Details

For the second stage classification, we have experimented with several classifiers (SVMs with several different kernels, Gradient Boosted Trees, MLPs) on a validation set, but found in nearly all circumstances the random forest performed similar or better than others. Trees in the RF were grown using the Gini criterion for splitting and in all situations we used 2000 estimators and the square root heuristic for the maximum number of features. The maximum depth was cross-validated using 8 folds. Data was balanced by drawing bootstrap samples with an equal class ratio. The posterior probability output by the RF was calculated as a mean of the estimated classes. The systems are trained using at most the ten most suspicious lesions per image found by the candidate detector, during testing no such threshold is applied to obtain highest sensitivity.

We implemented the network in Theano [16] and pointers provided by Bengio [13] were followed and very helpful. We used OxfordNet-like architectures [246] with 6 convolutional layers of $\{16, 32, 64, 128, 128\}$ with 3×3 kernels and 2×2 max-pooling on all but the fourth convolutional layer. A stride of 1 was used in all convolutions. Two fully connected layers of 300 each were added. An illustration of the network is provided in figure 6.4.

We employed Stochastic Gradient Descent (SGD) with RMSProp [55], an adaption of R-Prop for SGD with Nesterov momentum [257]. Drop-out [252] was used on the fully connected layers with $p = 0.5$. We used the MSRA [113] weight filler, a learning rate of 5×10^{-5} with a weight decay of 5×10^{-5}. To battle the strong class imbalance, positive samples

Table 3.2: Overview of results of individual feature sets along the 95% confidence interval (CI) obtained using 5000 bootstraps.

Feature group	**AUC**	**CI**
Candidate detector	0.858	[0.827, 0.887]
Contrast	0.787	[0.752, 0.817]
Texture	0.718	[0.681, 0.753]
Geometry	0.753	[0.721, 0.784]
Location	0.686	[0.651, 0.719]
Context	0.816	[0.781, 0.850]
Patient	0.651	[0.612, 0.688]
Equal Information	0.892	[0.864, 0.918]
All	0.906	[0.881, 0.929]

Table 3.3: Overview of results of the CNN combined with individual feature sets

Feature group added to CNN	**AUC**	**CI**
CNN Only	0.929	[0.897, 0.938]
Candidate detector	0.938	[0.919, 0.955]
Contrast	0.931	[0.91, 0.949]
Texture	0.933	[0.912, 0.950]
Geometry	0.928	[0.907, 0.946]
Location	0.933	[0.913, 0.950]
Context	0.934	[0.914, 0.952]
Patient	0.929	[0.908, 0.947]
All	0.941	[0.922, 0.958]

were presented multiple times during an epoch, keeping a $50/50$ positive/negative ratio in each minibatch. Alternatively, the loss function could be weighted, but we found this to perform worse, we suspect this is because rebalancing maintains a certain diversity in the minibatch. All hyperparameters were optimized on a validation set and the CNN was subsequently retrained on the full training set using the found parameters. All test patches were also augmented using the same augmentation scheme. On the validation set, this gave a small improvement. The best validation AUC was 0.90.

3.4.3 ROC Analysis

To first get an understanding of how well each feature set performs individually, we trained different RFs for each feature set and applied them separately to the test set. In all cases, the training procedure as described above was used. AUC values along with a 95% confidence interval, acquired using bootstrapping [69, 21] with 5000 bootstrap samples are shown in table 3.2.

The CNN was compared to the reference system with equal amount of information (i.e., excluding location, context and patient information) to get a fair performance comparison. Figure 3.6 shows a plot of the mean curves along with the 95% confidence interval obtained after bootstrapping. Results were not found to be significantly different $p = 0.2$ on the full ROC. Figure 3.7 shows a plot comparing the CNN with data augmentation to the network without data augmentation and with data augmentation and added manual features. Again bootstrapping was used to obtain significance. It is clear that the proposed data augmentation methods contributes greatly to the performance, which was also found to be significant ($p << 0.05$).

To combine the CNN with other descriptors, we extracted the features from the last fully connected layer and appended the other set (see figure 6.4). For each augmented patch, the additional features were simply duplicated. Table 3.3 shows results of the CNN combined with different feature sets, again with confidence interval acquired by bootstrapping with 5000 samples.

To investigate the degree to which a large dataset is really needed, we trained several networks on subsets, removing 40 percent of the malignant lesions. Results are provided in table 3.4. Since the differences are rather large, we did not perform significance testing. For all settings, we optimized the learning rate but kept all other hyperparameters equal to the ones found to be optimal for the full training set.

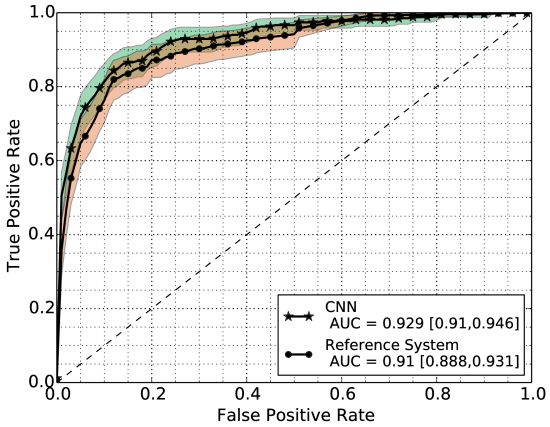


Figure 3.6: Comparison of the CNN with the reference system using equal information, i.e., only information represented in the patch used by the CNN, excluding context, location and patient information.

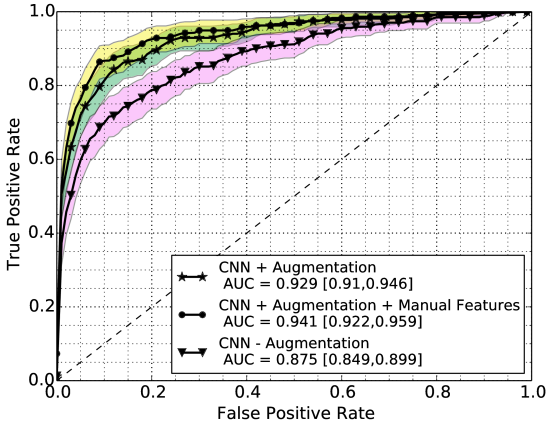


Figure 3.7: Comparison of the CNN without any augmentation, with augmentation and with added manual features.

Table 3.4: AUC values obtained when training the network on subsets of malignant lesions in the training set, keeping the same amount of normals.

Data Augmentation	60%	All
With	0.842	0.929
Without	0.685	0.875

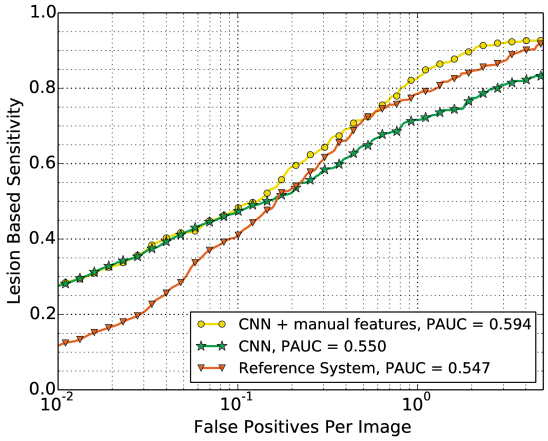


Figure 3.8: Lesion based FROC of the three systems. Please note that this concerns the full reference system, where context, location and patient features are incorporated

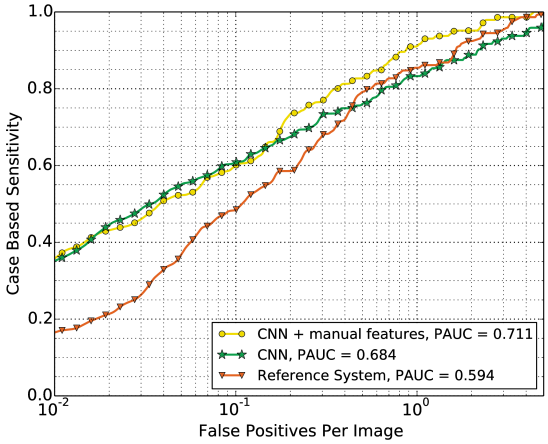


Figure 3.9: Case based FROC of the three systems. In areas of high specificity, the CNN and the addition of manual features is particularly useful. Please note that this concerns the full reference system, where context, location and patient features are incorporated

3.4.4 FROC Analysis

In practice, a CAD system should ideally be operating at a referral rate similar to that of a radiologists. To get a better understanding of the system's performance around this operating point, we compute the Partial Area Under the Curve (PAUC) on a log scale:

$$PAUC = \frac{1}{\ln[1] - \ln[0.01]} \int_{0.01}^{1} \frac{s(f)}{f} \mathrm{d}f \tag{3.2}$$

and generate Free Receiver Operator Characteristic (FROC) curves, to illustrate the numbers of false positives per image.

Plots of the FROCs of the full reference system (last line in table 3.2), the CNN only and the CNN plus manual features are shown in figures 3.8 and 3.9. To further investigate which features are helpful at high specificity, we compute PAUC for each feature set individually. Results are shown in table 3.5. We see a significant difference comparing the CNN with additional features to the reference system $P = 0.015$ on a lesion level and $P = 0.0002$ on a case level.

3.4.5 Human Performance

In previous work in our group, performance of the CAD system was compared to the performance of a radiologists at an exam level, a collection of four images, which contains more information than only a patch, such as context in the mammogram, symmetrical difference between two breast, the relation between the CC and MLO views. To get a better

Table 3.5: Partial Area under the FROC of different systems. P-values are referring to the comparison between the CNN with additional features and the CNN without the specific feature group. In this case, the reference system is the *full* system, including context, location and patient information.

	Lesion	Case	P, lesion	P, case
CNN	0.550	0.684	1	1
Reference System	0.547	0.594	0.451	**0.013**
CNN + candidate det.	0.590	0.701	<**0.0001**	**0.026**
CNN + contrast	0.571	0.704	**0.011**	0.0758
CNN + texture	0.574	0.705	**0.0062**	0.067
CNN + topology	0.561	0.700	**0.0286**	0.132
CNN + location	0.576	0.707	**0.0038**	0.0516
CNN + context	0.578	0.700	**0.0028**	0.121
CNN + patient	0.576	0.704	**0.0034**	0.0784
CNN + all features	0.594	0.711	<**0.001**	**0.04**

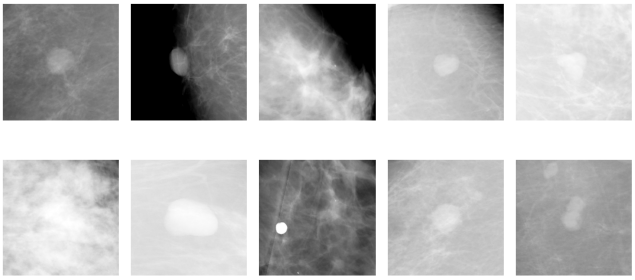


Figure 3.10: Top misclassified negatives by the CNN. The second sample in the first row is simply the nipple and the third sample in the second row displays fat necrosis. Both are obviously normal patches and are filtered out using additional feature sets.

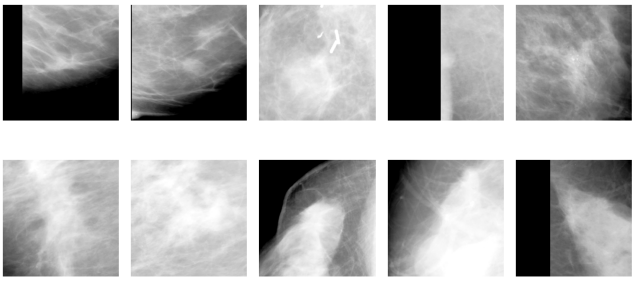


Figure 3.11: Top misclassified positives by the CNN, most samples are very large lesion unlikely to be found in the screening population and therefore underrepresented in the training set.

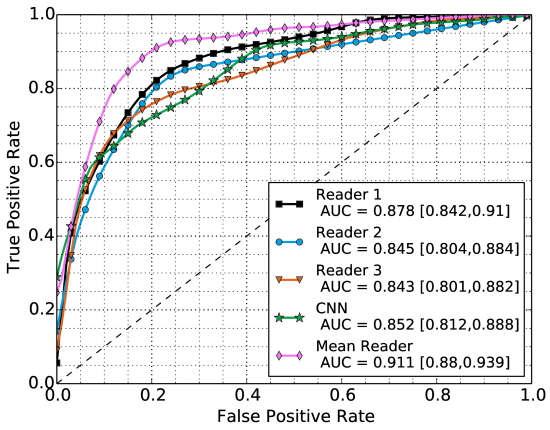


Figure 3.12: Comparison between the CNN and three experienced readers on a patch level.

understanding of how close the CNN is to human performance on a patch level and how much more room there is for improvement in this sub part of the pipeline, we performed a study where we measured the performance of experienced readers on a patch level, providing the reader with the same information as the CNN. The group of readers consisted of one experienced reader (non-radiologist) and two experienced certified radiologists. To get an idea of the performance that can at least be obtained on this set, the mean of the three readers was also computed by simply averaging the scores that each of the three readers assigned to each patch.

Patches were extracted from the mammogram processed by the manufacturer for optimal viewing and were shown at a normal computer screen at a resolution of 200 micron. Microcalcifications are difficult to see in this setting, but all structures relevant for soft tissue lesions are intact and readers did not report difficulties. The readers were provided with a slider and instructed to score the patch between zero and one hundred based on their assessment of the suspiciousness of the patch.

As a test set, we used all masses that were used in Hupse et al. [126] and selected an equal amount of negatives, that were considered the most difficult by the candidate detector, resulting in 398 patches. This gives a representative set of difficult samples and allows for larger differences between readers and the CNN, but is biased towards a set difficult for the reference system, which was therefore left out of the comparison (obtained AUC was 0.64 on this set). Figure 3.12 shows the ROC curves resulting from the reader study. Again, to test significance we used bootstrapping and two sided testing to get a significance score. We found no significant difference between the CNN and any of the readers: CNN vs reader 1: $p = 0.1008$, CNN vs reader 2: $p = 0.6136$, CNN vs reader 3: $p = 0.64$, but found a significant difference between the CNN and the mean of the human readers ($p = 0.001$).

3.5 Discussion

To get more insight into the performance of the network, examples of the top misclassified positives and negatives are shown in figure 3.11 and 3.10 respectively. A large part of the patches determined as suspicious by the network are benign abnormalities such as cysts and fibroadenomae or normal structures such as lymph nodes or fat necrosis. Cysts and lymph nodes can look relatively similar to masses. These strong false positives occur due to the absence of benign lesions in our training set. In the future we plan to add these to the training set and perform three-class classification or train a separate network to discriminate these lesions properly.

The majority of 'misclassified' positives are lesions ill-represented in the training data, either very subtle or extremely large. When using CAD as a second reader, these will not influence the referral decision much, as they are clearly visible to a human, but when using the computer as an independent reader, these issues need to be solved. In preliminary experiments, we have seen that many of these misclassifications can be prevented by considering the contralateral breast and plan to work on this in the future.

From the results in tables 3.3 and 3.2 we can see that individually, apart from the candidate detector, contrast and context are useful features. Although age and screening round are some of the most important risk factors, we do not see clear improvements when added as features, which is slightly disappointing. To get training data, we took negative patches only from normal images, but not only from normal exams, to get as many data points as possible. A possible

explanation for the disappointing performance may be that the relation between age and cancer is more difficult to learn in the setting, since it is a relation that exist on an exam level.

To add features, we have used a second classification stage. This has the advantage it is easy to evaluate which features add information, without retraining a network and re-optimizing the parameters, which can take several weeks to do properly. On top of this, the learned feature representation of the CNN is the same in all situations, rendering comparison more reliable. A major disadvantage, however, is that the training procedure is rather complicated. Other more elegant methods such as coding features as a second channel, as done by [180] or adding the features in one of the fully connected layers of the network during training could be better strategies and we plan to explore this in future work.

We have made use of a more shallow and scaled down version of the networks proposed by Simonyan et al. [246], who obtain best performance on ImageNet with a 19 layer architecture with four times the amount of kernels in each layer. In initial experiments, we have worked with Alexnet-like architectures, which performed worse on our problem, obtaining and AUC of around 0.85 on the validation set. We have also experimented with deeper networks and increasing the amount of kernels, but found no significant improvement on the validation set (0.896 vs 0.897 of the network with larger capacity and 0.90 of 9 layer network). We suspect that with more data, larger capacity networks can become beneficial. The problem could be less complex than classifying natural images since it concerns a two-class classification in the current setting and we are dealing with gray scale images, contrary to the thousands of classes and RGB data in ImageNet [220]. Therefore, more shallow and lower capacity networks than the one found optimal for natural images could suffice for this particular problem.

In our work, we made extensive use of data augmentation in the form of simply geometric transformations. We have also experimented with full rotation, but this creates lesions not expected during testing, due to the zero padding. This could be prevented using test time augmentation, but when used in a sliding window fashion this is not convenient. The ROC curves in figure 3.7 show a clear increase in performance for the full dataset. The results in table 3.4 show the current data augmentation scheme improves performance for large amounts of data but not for small amounts of data. We suspect in the latter setting, the network overfits and more regularization is needed. These results may be different when fully optimizing the architecture and augmentation procedure for each setting individually. More research is needed to draw clear conclusions. However effective, data augmentation is a rather computationally costly procedure. A more elegant approach would be to add the invariance properties in the network architecture, which is currently being investigated in several papers [96, 131]. On top of the geometric transforms, occluding tissue is an important source of variance, which is more challenging to explicitly code in the network architecture. In future work, we plan to explore simulation methods for this.

In this work, we have employed a previously developed candidate detector. This has two main advantages: (1) it is fast and accurate (2) the comparison with the traditional CAD system is straightforward and fair, since exactly the same candidate locations are trained with and evaluated on. The main disadvantage is that the sensitivity is not hundred percent, which causes lesions to be missed, although the case-based performance is close to optimal. In future work, we plan to explore other methods, such as the strategy put forth by Ciresan et al. [47], to train the system end-to-end. This will make training and classification less cumbersome and has the potential to increase the sensitivity of the system.

In this work we have compared the CNN to a state-of-the art CAD system [126], which was combined with several other features commonly used in the mammography CAD literature. A random forest was subsequently used, that performs feature selection during its training stage. We think the feature set we used is sufficiently exhaustive to include most features commonly used in literature and therefore think similar conclusions hold for other state-of-the art CAD systems. To the best of our knowledge, the Digital Database of Screening Mammography (DDSM) is the only publicly available dataset, which comprises of digitized screen film mammograms. Since almost all screening centers have migrated to digital mammography, we have elected not to run our system on this dataset, because we think the clinical relevance is arguable. On top of this, since this entails a *transfer learning* problem, the system may require retraining to adapt to the older modality.

The reader study illustrates the network is not far from the radiologists performance, but still substantially below the mean of the readers, suggesting a large performance increase is still possible. We suspect that some other augmentation methods as discussed above could push the network a bit further, but expect more training data, when it becomes available will be the most important factor. Also, we feel still employing some handcrafted features that specifically target weaknesses of the CNN may be a good strategy and may be more pragmatic and effective than adding thousands of extra samples to the training set.

3.6 Conclusion

In this chapter we have shown that a deep learning model in the form of a Convolutional Neural Network (CNN) trained on a large dataset of mammographic lesions outperforms a state-of-the art system in Computer Aided Detection (CAD) and therefore has great potential to advance the field of research. A major advantage is that the CNN learns from data and does not rely on domain experts, making development easier and faster. We have shown that the addition of location information and context can easily be added to the network and that several manually designed features can give some small improvements, mostly in the form of 'common sense': obviously false negatives will no longer be considered as such. On top of this, we have compared the CNN to a group of three experienced readers on a patch level, two of which were certified radiologist and have show that the human readers and CNN have similar performance.

Chapter 4

Invariant features for discriminating cysts from solid lesions in mammography

Appeared in
Invariant features for discriminating cysts from solid lesions in mammography. - *Thijs Kooi and Nico Karssemeijer*. International Workshop on Digital Mammography. Springer International Publishing, 2014.

Abstract

Feature extraction is an integral part of all Computer Aided Diagnosis (CAD) systems. Due to the presence of fibroglandular tissue however, measurements are perturbed by unwanted influences and therefore, the same descriptor will yield different values for different amounts of occluding structures. To aid the statistical learning used for classification, we need to design features that are *invariant* to unwanted influences. In this paper, we propose a simple model of the tumour and its surrounding tissue and show how this model can be used to derive descriptors that are invariant to obscuring tissue, rather than heuristically defining a set of descriptors, which is common practice in many CAD papers. We tailor the descriptors to optimally discriminate between tumours and cysts, by assuming a parametric form of the lesions. Results show a significant discriminative improvement over simple, more commonly used contrast features and we obtained an AUC of 0.77 using both CC and MLO images.

4.1 Introduction

Many CAD systems operate in a two-stage fashion where in the first stage, candidate lesions are detected on a pixel level and in a second stage, the lesions are segmented and new region-based features are computed. The second feature set is subsequently fed to a statistical learning machine, which is expected to give a more accurate estimate of the lesion's nature. The contrast, i.e., the relation between the segmented lesion and its surrounding tissue can tell us something about the disease and is often used as a feature in this stage. Due to the presence of fibroglandular tissue however, measurements are perturbed by unwanted influences and therefore, the same descriptor will yield different values for different amounts of occluding tissue. Given enough training data, we could expect the classification machine to generalise to a sufficient extent. Unfortunately, data is still scarce and therefore we need to aid the learner and design features that are *invariant* [98, 123] to unwanted influences, yet covary with the factor we are interested in, i.e., we want descriptors that yield the same value regardless of obscuring structures, yet reliably characterise the nature of the lesion.

In the first part of this chapter, we propose a simple model of the tumour and its surrounding tissue by making some assumptions on the tumour growth and segmentation and show how to this model can be used to derive descriptors that are invariant to unwanted influences, rather than heuristically defining a set of descriptors, which is common practice in many CAD papers. In the second part, we tailor the descriptors to optimally discriminate between different geometrical entities, by assuming a parametric form. We apply the descriptors to the problem of discriminating between tumours and cysts: benign fluid-filled sacks exhibiting similar image characteristics to tumours in a mammogram. Experiments are done on simulated lesions, placed in a standard mammography background and on a database of clinical cases with a representative sample of cysts and masses. Results on both sets show a significant discriminative improvement over simple, more commonly used contrast features. Our method relies on the fact that the observed shape of the lesion on the image plane, is the result of a summation of attenuation along the z-axis, which we will henceforth refer to as the *z-integral* and that different geometrical structures will expose different z-integrals. Even though we can not observe the exact spatial structure of the body, we can get an idea of its shape, by looking at the moments of the distribution of the structure.

The rest of this chapter is organised as follows. In section 4.2, we will describe our model and preprocessing methods, followed by a brief derivation of the features in section 4.3 and details on the normalisation in section 4.4. We will present our experiments and discuss results in section 4.5, followed by a conclusion in section 4.6.

4.2 Lesion Model

In order to derive invariant descriptors, we first need a proper definition of invariance. We will call a descriptor $\mathcal{D}$ of some signal s invariant to a transformation T if it holds that:

$$D(T(s))) = D(s)$$

In our setting this means that we want to find a descriptor of a tumour (s) is such a way that if the same tumour is found in two different women with different amount of occluding tissue (T), the descriptor will give the same value. A trivial way to make a descriptor invariant is to simply assign 1 to every exposition of the signal. This however, obliterates all discriminative power and we should therefore aim for an optimum in the trade-off between descriptiveness and invariance.

The first step in our method is to compute a *dense tissue map* of the image [275], which is acquired be means of a physics based image model derived in previous work. We assume the breast is composed of dense and fatty tissue, with corresponding attenuation coefficients. Using empirical data from literature, we get an estimate of the amount of dense and fatty tissue at each pixel, where the former is used in our method. To subsequently derive our descriptors, we propose the following simple model of the lesion f:

$$I(x,y) = \begin{cases} F_z(x,y) + \epsilon & \text{if } (x,y) \text{ lies inside the 2D projection of } f \\ \epsilon & \text{else} \end{cases} \tag{4.1}$$

where $I(x,y)$ indicates the image value at location (x,y), $\epsilon \sim P(\epsilon)$ is the integral along the z-axis of the nuisance term we are trying to ignore, coming from some undefined distribution and $F_z(x,y) = \int f(x,y,z)dz$ the z-integral of the lesion f we are trying to describe. The z-axis is chosen to be parallel to the direction of x-ray quanta. This model assumes that the 2D segmentation of the projected lesion is correct and that the tumour grows in such a way that the distribution of tissue in the surrounding region is the same as above and below it. An illustration is provided in figure 4.1. Using this model, we can not infer anything about the exact spatial layout of f, due to the noise term ϵ. However, by

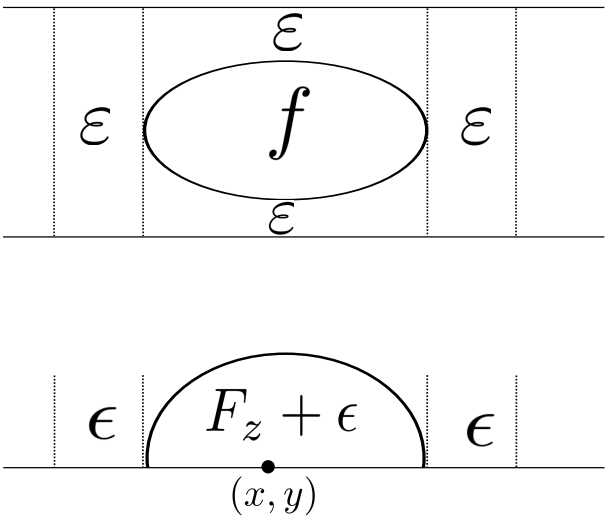


Figure 4.1: Illustration of a lateral view of a lesion and its z-integral on the image plane. Here ε indicates the tissue in the breast and ϵ the integral of ε appearing on the image.

looking at the descriptive statistics of the values inside the segmented region ($F_z + \epsilon$) and the values in the surroundings of the region (ϵ), we can still derive descriptors of its shape that are invariant to nuisance ϵ.

4.3 Moment Invariants

Under the definition of invariance and lesion model we proposed, a descriptor of the mean, invariant to nuisance signal ϵ is given by:

$$\mathbb{E}[F_z] = \mathbb{E}[F_z + \epsilon] - \mathbb{E}[\epsilon] \tag{4.2}$$

Similar to the mean, the variance will give us an indication of the shape of the z-integral. The sum of variance to two correlated random variables is given by:

$$Var[F_z + \epsilon] = Var[F_z] + Var[\epsilon] + 2Cov[F_z, \epsilon] \tag{4.3}$$

We can observe $Var[F_z + \epsilon]$ and $Var[\epsilon]$ in our image and are interested in $Var[F_z]$. Unfortunately, we can not observe $Cov[F_z, \epsilon]$. However, by looking at the covariance of $F_z + \epsilon$ and ϵ, we can find an expression for $Var[F_z]$. We can show that the covariance of two correlated random variables X and Y can be written as (a proof of this is left out for brevity)

$$Cov[X, Y] = Cov[X + Y, Y] - Var[Y]$$

Plugging this into equation (4.3) and rearranging terms, we have that

$$Var[F_z] = Var[F_z + \epsilon] + Var[\epsilon] - 2Cov[F_z + \epsilon, \epsilon] \tag{4.4}$$

These two features are general descriptors of the z-integral of any geometric body, though in not all situations they may provide useful, discriminative information. We will now consider the problem of discriminating tumours from cysts, by assuming a parametric form and tailor the descriptors to optimally differentiate between these entities. Most tumours are relatively hard, solid objects that are not easily compressible, whereas cysts are softer and therefore more likely to change shape. If we assume both are initially spherical, we can expect the tumour to retain shape, but the cyst to transform to an ellipsoidal form, due to the compression of the breast in the recording of the mammogram. Under these assumptions, we can further refine the descriptor and normalise for scale.

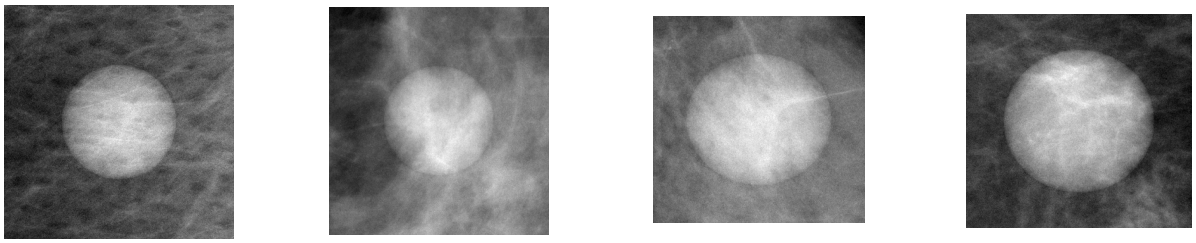


Figure 4.2: Illustration of simulated lesions. The left two represent masses and the right two cysts.

4.4 Scale Normalisation

Simply increasing the size of a spherical body, will yield a varying feature response. This makes it possible for ellipsoidal and spherical objects to reach similar descriptors, in spite of their apparent disparity. We therefore want to normalise with respect to scale in order to discriminate properly between the two. To this end, we propose the following normalisations. By looking at the z-integral of a sphere

$$F_z = 2\sqrt{\left(r^2 - x^2 - y^2\right)}$$

we can see that the expected value as a function of radius r is given by:

$$\mathbb{E}[F_z; r] = \frac{1}{A(r)} \int_{-r}^{r} \int_{-r}^{r} 2\sqrt{\left(r^2 - x^2 - y^2\right)} dx\, dy$$

where $A(r)$ indicates the area of the projection of the sphere as a function of its radius. The integral is simply the volume of the sphere, so therefore:

$$\mathbb{E}[F_z; r] = \frac{V(r)}{A(r)} = \frac{1}{\pi r^2} \frac{4}{3} \pi r^3 = r \frac{4}{3}$$

Normalisation by r evidently results in a constant. Using a similar procedure for the second moment descriptor, we will find that a normalisation by r^2 will yield a constant value. The derivation is left out here for brevity. The scale normalised descriptors $\hat{\mathbb{E}}[F_z]$ and $\hat{Var}[F_z]$ are now given by:

$$\hat{\mathbb{E}}[F_z] = \frac{\mathbb{E}[F_z + \epsilon] - \mathbb{E}[\epsilon]}{r} \tag{4.5}$$

and

$$\hat{Var}[F_z] = \frac{Var[F_z + \epsilon] + Var[\epsilon] - 2Cov[F_z + \epsilon, \epsilon]}{r^2} \tag{4.6}$$

The following section will present our experimental setup and results acquired when applying this method.

4.5 Experiments

The experiments are set up to show that each normalisation and both the first and second moment are contributors to the discriminative power of the features. In a first test, we investigated how our model holds in a more or less ideal situation. This way we can see to what extent the assumptions in our model are violated when subsequently testing on real data. We placed masses (z-integrals of spheres) and cysts (z-integrals of ellipsoidal shapes) at random in a mammography background. The parameters of the spheres and ellipsoids were varied at random and the mix between z-integral and mammographic background was also varied randomly, resulting in some clear lesion and some very subtle ones. Examples of simulated lesions are given in 4.2. In a second test, we applied the methods to real data. Regions were segmented using a dynamic programming algorithm, that has previously been shown to be successful for this task [270]. The data was collected locally from symptomatic women and high risk screening. We removed lesions that were on or close to the pectoralis, because our density algorithm does not support reliable density estimates on the pectoralis yet. A

Table 4.1: Comparison of different feature sets in simulation

Estimate of mean	Estimate of Variance	AUC	P-value against normalised features
$\mathbb{E}[F_z+\epsilon]-\mathbb{E}[\epsilon]$	$Var[F_z+\epsilon]$	0.55 ± 0.01	$\ll 0.001$
$\mathbb{E}[F_z+\epsilon]-\mathbb{E}[\epsilon]$	$Var[F_z+\epsilon]-Var[\epsilon]$	0.58 ± 0.02	$\ll 0.001$
$\frac{\mathbb{E}[F_z+\epsilon]-\mathbb{E}[\epsilon]}{r}$	$Var[F_z+\epsilon]-Var[\epsilon]$	0.86 ± 0.07	$\ll 0.001$
$\frac{\mathbb{E}[F_z+\epsilon]-\mathbb{E}[\epsilon]}{r}$	$\frac{Var[F_z+\epsilon]-Var[\epsilon]}{r^2}$	0.93 ± 0.00	$\ll 0.001$

Table 4.2: Comparison of different feature sets on our dataset

Estimate of mean	Estimate of Variance	AUC	P-value against normalised features
$\mathbb{E}[F_z+\epsilon]-\mathbb{E}[\epsilon]$	$Var[F_z+\epsilon]$	0.52 ± 0.07	$\ll 0.001$
$\mathbb{E}[F_z+\epsilon]-\mathbb{E}[\epsilon]$	$Var[F_z+\epsilon]-Var[\epsilon]$	0.54 ± 0.01	$\ll 0.001$
$\frac{\mathbb{E}[F_z+\epsilon]-\mathbb{E}[\epsilon]}{r}$	$Var[F_z+\epsilon]-Var[\epsilon]$	0.63 ± 0.08	0.006
$\frac{\mathbb{E}[F_z+\epsilon]-\mathbb{E}[\epsilon]}{r}$	$\frac{Var[F_z+\epsilon]-Var[\epsilon]}{r^2}$	0.62 ± 0.07	0.019

similar thing was done for lesion on or close to the image edge, because we can not reasonably assume to get accurate contrast information from here. Future work will revolve around finding new methods for this. During annotation, all lesions were given a subtlety score by an independent annotator and we remove lesions with extreme subtlety. This left 94 cysts and 173 masses. Images were recorded using a GE mammography machine. Samples were classified by means of a linear logistic regression model, trained using iteratively reweighted least-squares (IRLS). To estimate test performance, we split the train and test data 100 times, resulting in 100 estimates of the ROC and AUC, over which we compute analytical statistics.

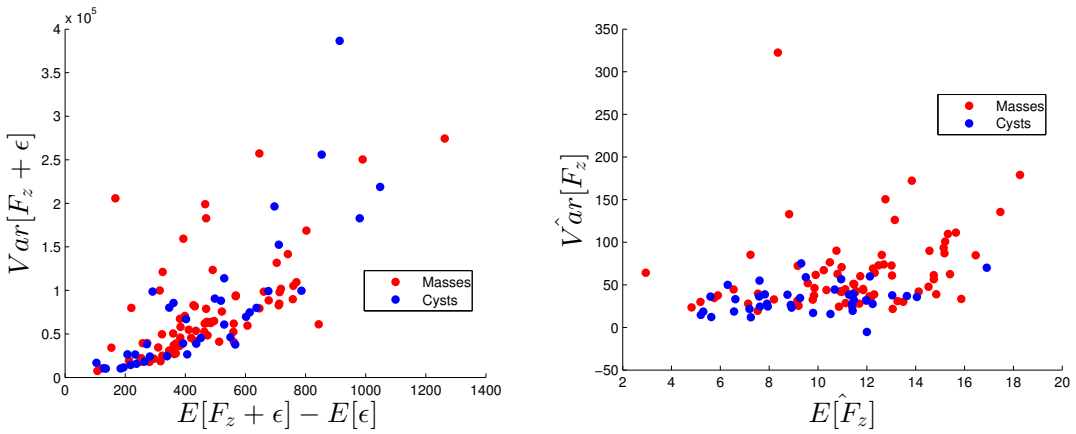
4.5.1 Implementation details

The computation of covariance assumes an ordered set of pairs and therefore, the use of covariance in (4.6) requires some elaboration. Considering the fact that the noise we are interested in (the occluding tissue) has some spatial coherence, i.e., the noise between surrounding pixels is strongly correlated, it seems reasonable to assume that points on each side of the acquired segmentation boundary will have roughly the same noise value. Using this intuition, we construct a set of points by taking values along the fitting region on each side of the border and compute the covariance from this set.

The radius parameter we have described before, assumes a circular fitting region, which is in practice never the case. We therefore take the radius as half the maximum diameter of the fitted region. Lastly, the size of the surrounding border (ϵ) is chosen in such a way that the amount of pixels in the lesion and surrounding area are approximately similar if the particular image allows this.

4.5.2 Results

On simulated data, our normalised features as described in (4.5) and (4.6) gave an average $AUC = 0.95 \pm 0.004$. Table 4.1 shows the results of a comparison between these and several feature sets that are increasingly similar to our descriptors, thereby proving the value of each step in our methods. The first and second column show the estimate of the first and second moment, the third column the acquired average AUC and the fourth column the P-value tested against our reference features. Significance estimates were acquired by means of a Kruskal-Wallis rank test. On real data, our normalised features as described in (4.5) and (4.6) obtained an average AUC of 0.65 +/- 0.07. Table 4.2 shows results of to several other feature sets, in increasing complexity. Again, P-values were computed using a Kruskal-Wallis test. In a third experiment, we averaged the feature output of lesions in CC and MLO images in order to reduce measuring errors. In this setting, our normalised features obtained an AUC of 0.77 ± 0.09, compared to an AUC of 0.66 ± 0.12 using a CC/MLO average of the feature described in the last row of table 1. This was found significant using a Kruskal-Wallis test ($p \ll 0.001$). Ilustrations of the final feature spaces of the reference and normalised features are given in figure 4.3.



(a) Reference Features

(b) Normalised Features

Figure 4.3: Illustration of the Feature Space for Real Data, Averaged between CC and MLO

4.5.3 Discussion

In an ideal setting, we can see our descriptors fare well and vastly outperform the simple, commonly applied descriptors, described in the first row of table 4.1. Even though the classification performance on real data is still relatively poor using a single image, we can clearly see an improvement using the proposed descriptors, as is seen from table 4.2. The progression to more complex descriptors as is presented in the tables, suggests that both a normalisation with respect to the surrounding region and a normalisation with respect to scale are contributors to the improvement and that not only the proposed estimate of the mean, but also the estimate of variance yields useful discriminative information. This can also be seen from the plot of the feature space in figure 4.3, where two clearer clusters appear in the right image.

To make the estimates less susceptible to outliers, we have tried to replace the regular estimates of location and scale with robust estimates. We tried using the median and winsorising to replace the standard estimate of location and we used the inter quartile range, median absolute difference and Minimum Discriminant Covariance algorithm to generate a robust estimate of the (co)variance. In the case of location, this yielded no difference and in the case of scale, this performed significantly worse. Apart from applying the methods in the density image, we have also tried our method on raw FFDM data, but both methods proved substantially worse in this setting, reaching a classification performance only slightly above what one would expect from a random assignment of labels.

In our dataset, we found a difference in compresssion forces between the CC and MLO image. (Average compression force CC: 160.9+/-33.94 and for MLO: 169.0+/-29.6), which was found significant by a two-sided t-test ($p = 0.008$). Using the same intuition that cysts are more strongly compressend than solid lesions, one would expect a difference in feature response between CC and MLO for cysts and that this difference grows stronger with discrepancy in compression. Based on this intuition, we attemted to normalise feature difference based on a difference in compression force. This, however, seriously decreased performance. We suspect there to be too much noise in the system for this to work properly.

4.6 Conclusion

In this chapter, we showed that by assuming a simple model of a tumour and its surrounding area, lesion features can be derived that in the first place are invariant to tissue occluding the lesion and by assuming a parametric form of the lesion, invariant to scale. We showed our model-derived features outperform several simple, heuristically chosen features, typically applied in other systems. This is our first step on deriving model based descriptors. The model we use, assumes the tumour grows in a specific, highly simplified way and has a simplified shape. In the data we observed so far, the growth seems to hold. However the tumour in our database seem to have a very sharp gradient around the edges, but the centre of the lesion appears flattened. In future work, it may be worthwhile to investigate different parametric forms in the derivation of features. It can be shown that for spherical and ellipsoidal z-integrals, the skewness and kurtosis is exactly the same (and invariant to scale) and therefore for our test problem not useful. For other structures however, these can still prove useful, although derivation is quite cumbersome.

Derivation of Second Moment

We would like to show that:

$$Cov[X+Y,Y] = Cov[X,Y] + Var[Y]$$

By definition of covariance, we have that:

$$Cov[X+Y,Y] = \frac{1}{N}\sum_{n=1}^{N}\Big((X+Y)Y - (X+Y)\mathbb{E}[Y] - Y\mathbb{E}[X+Y] + \mathbb{E}[Y]\mathbb{E}[X+Y]\Big)$$

where we have dropped the indices for brevity. Isolating the summations, we find that:

$$= \frac{1}{N}\sum_{n=1}^{N} XY + Y^2 - \frac{1}{N}\sum_{n=1}^{N}(X+Y)\mathbb{E}[Y] - \frac{1}{N}\sum_{n=1}^{N} Y\mathbb{E}[X+Y] + \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[Y]\mathbb{E}[X+Y]$$

Because by definition $\mathbb{E}[X] = \frac{1}{N}\sum_{n=1}^{N} X_n$ for some random variable X and $\mathbb{E}[a] = \frac{1}{N}\sum_{n=1}^{N} a = a$ for some constant a, we can write this as:

$$= \mathbb{E}[XY] + \mathbb{E}[Y^2] - \mathbb{E}[Y]\mathbb{E}[X+Y] - \mathbb{E}[Y]\mathbb{E}[X+Y] + \mathbb{E}[Y]\mathbb{E}[X+Y]$$

The forth and the fifth term cancel out. Using the identity $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$, we see that

$$= \mathbb{E}[XY] + \mathbb{E}[Y^2] - \mathbb{E}[Y]\mathbb{E}[X] - \mathbb{E}[Y]^2$$

since we know that $Var[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$, this is equal to

$$= \mathbb{E}[XY] - \mathbb{E}[Y]\mathbb{E}[X] + Var[Y]$$

and therefore

$$= Cov[X,Y] + Var[Y]$$

Chapter 5

Discriminating solitary cysts from solid lesion in mammography using a deep convolutional neural network

Appeared in:
Discriminating solitary cysts from soft tissue lesions in mammography using a pretrained deep convolutional neural network. - *Thijs Kooi, Bram van Ginneken, Ard den Heeten and Nico Karssemeijer*. Medical physics 44.3 (2017): 1017-1027.

Abstract

Purpose It is estimated that 7% of women in the western world will develop palpable breast cysts in their lifetime. Even though cysts have been correlated with risk of developing breast cancer, many of them are benign and do not require follow-up. We develop a method to discriminate benign solitary cysts from malignant masses in digital mammography. We think a system like this can have merit in the clinic as a decision aid or complementary to specialised modalities.

Methods We employ a deep Convolutional Neural Network (CNN) to classify cyst and mass patches. Deep CNNs have been shown to be powerful classifiers, but need a large amount of training data which for medical problems is often difficult to come by. The key contribution of this chapter is that we show good performance can be obtained on a small dataset by pretraining the network on a large dataset of a related task. We subsequently investigate the following: (1) when a mammographic exam is performed, two different views of the same breast are recorded. We investigate the merit of combining the output of the classifier from these two views. (2) We evaluate the importance of the resolution of the patches fed to the network. (3) A method dubbed tissue augmentation is subsequently employed, where we extract normal tissue from normal patches and superimpose this onto the actual samples aiming for a classifier invariant to occluding tissue. (4) We combine the representation extracted using the deep CNN with our previously developed features.

Results We show that using the proposed deep learning method, an Area Under the ROC Curve (AUC) value of 0.80 can be obtained on a set of benign solitary cysts and malignant mass findings recalled in screening. We find that it works significantly better than our previously developed approach by comparing the AUC of the ROC using bootstrapping. By combining views, the results can be further improved, though this difference was not found to be significant. We find no significant difference between using a resolution of 100 versus 200 micron. The proposed tissue augmentations give a small improvement in performance, but this improvement was also not found to be significant. The final system obtained an AUC of 0.80 95% with confidence interval $[0.78, 0.83]$, calculated using bootstrapping. The system works best for lesions larger than 20mm where it obtains an AUC value of 0.87.

Conclusion We have presented a Computer Aided Diagnosis (CADx) method to discriminate cysts from solid lesion in mammography using features from a deep Convolutional Neural Network (CNN) trained on a large set of mass candidates, obtaining an AUC of 0.80 on a set of diagnostic exams recalled from screening. We believe the system shows great potential and comes close to the performance of recently developed spectral mammography. We think the system can be further improved when more data and computational power becomes available.

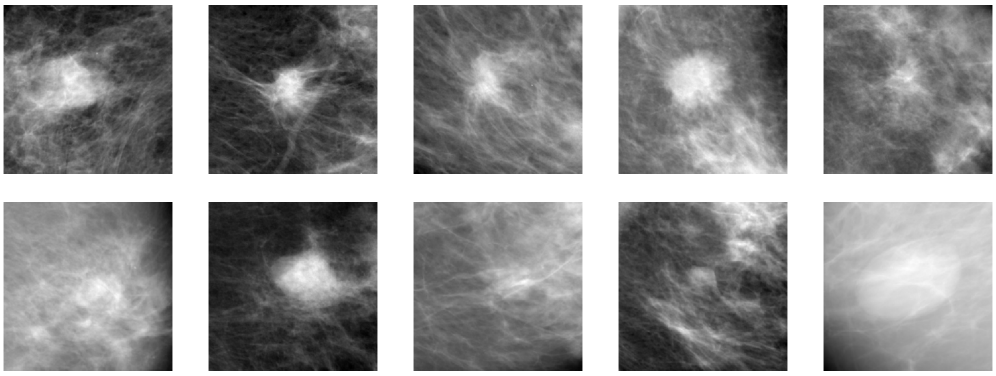


Figure 5.1: Typical examples of malignant masses (top row) and benign solitary cysts (bottom row) in our dataset. Each patch captures $5.2cm$ or 260 pixels at a resolution of 200 micron. Cysts are fluid filled sacs and part of normal tissue and malignant masses are signs of cancer and require follow up examination. Unfortunately, they are often difficult to distinguish during mammography screening meaning cancers can be missed or women are recalled unnecessarily for a follow up exam. We are developing a Computer Aided Diagnosis (CADx) method to discriminate malignant masses from benign cysts in mammography.

5.1 Introduction

It is estimated that 7% of women in the western world will develop palpable breast cysts in their lifetime. Even though cysts have been correlated with risk of developing breast cancer [62], many of them are benign and do not require follow-up. On mammography, benign cysts and solid lesion can be difficult to discriminate and consequently, many women are being recalled unnecessarily for a second diagnostic exam or core needle biopsy. Literature suggest 20% [72] to 37% [240] of recalls can be attributed to benign solitary cysts. False positives have been shown to cause severe psychological stress up to three years after diagnosis, sometimes as high as women diagnosed with breast cancer [26].

To better differentiate between findings, ultrasound is commonly used in diagnostic exams, but this is typically not available during screening. In a recent study, Erhard et al. [72] employed spectral mammography, an extension of mammography utilising two energy thresholds with a single exposure, which allows measuring attenuation and consequently cystic and lesion volume accurately. Two features are computed, the first based on the ratio of cystic against lesion volume and a second one measuring the cystic diameter. A Linear Discriminant Analysis (LDA) is subsequently used for classification. Their method obtained an Area under the ROC Curve (AUC) of 0.88 with a median specificity of 56% at 99% sensitivity. The authors estimate the employment of their method in the clinic would result in a reduction of 32% of recalls based on well-defined solitary lesions and an overall reduction in recall of 6% [72].

In this chapter, we present a CADx system to discriminate benign cysts from solid lesion in regular digital mammography. The key contribution is showing that a deep Convolutional Neural Network (CNN) can be used effectively on a small data set by pretraining the network on other data. Contrary to related work [236, 263], the network is not pretrained on natural images, but on a large dataset of soft tissue lesion candidates from a set of screening mammograms. Features are extracted from the network and a second classifier is trained for the diagnosis task. The method is compared to the only previously developed method for this problem, where a model of the cysts and solid lesions was applied to derive two descriptors invariant to surrounding tissue.

During a mammographic exam two views of each breast are typically recorded and when radiologists evaluate a case, the information in both is taken into account. We show that by combining these two views, better classification performance can be obtained. We subsequently show that by augmenting samples with normal tissue from normal areas in the breast higher performance can be obtained. Lastly, we combine the features extracted from the CNN with the previously developed method to obtain our final results.

The method is evaluated on a large dataset of roughly 1000 malignant masses and 600 cysts. Typical examples of masses and cysts in our dataset are shown in figure 5.1. To the best of our knowledge, this is the best performance

Table 5.1: Overview of dataset. We used two different sets: (1) A large dataset of screening mammograms containing normal exams and biopsy proven malignant masses (i.e., no cysts or other benign abnormalities) that was used to learn the features from the CNN on and (2) a set of diagnostic exams referred from screening containing biopsy proven malignant masses and benign solitary cysts. For the first dataset, we used a separate validation set to optimise the CNN's hyperparameters on, the amount of training and validation samples are indicated as (# training/ #validation). Since the second dataset is relatively small, we used nested cross-validation and hence no fixed validation or test set was employed.

Screening Exams (Hologic)	**Images**	**Exams**
Normal	(73102/21913)	(20000/8979)
Malignant masses	(1487/342)	(1000/205)
Diagnostic Exams (GE)		
Malignant masses	1108	586
Solitary Cysts	696	370
Total	1804	956

reported in literature for this problem in digital mammography and the first method that investigates transfer between related medical tasks in deep CNNs.

The rest of this chapter is divided in 5 sections. In section 5.2, we will provide an overview of the data. Section 5.3 will give details on the CNN setup and how data augmentation is applied, followed by a brief description of the previously published method which we compare the CNN against. In section 5.4, we provide experimental details and results, followed by a conclusion in section 5.5 and discussion in section 5.6.

5.2 Data

We make use of two different datasets. The first dataset concerns a large collection of screening mammograms obtained from a screening program in The Netherlands (*bevolkingsonderzoek midden-west*) and recorded using a Hologic Selenia digital mammography system at a resolution of 70 micron. All tumours are biopsy proven malignancies and annotated under the supervision of a certified radiologist. The data was randomly split into a training and validation set on a patient level. This dataset is used to pretrain the CNN.

The second set was used for retraining and producing the final results, which were obtained using nested cross-validation with 8 inner and outer folds with all data split on a patient level. It consists of diagnostic exams of women who were recalled in screening for a suspicious mass lesion. Again, all malignant lesions are biopsy proven. Images were recorded using a GE Senograph 2000D(S) at an original resolution of 100 micron. All soft tissue lesions used in this study were either marked as ductal carcinoma in situ, invasive ductal carcinoma or invasive lobular carcinoma. Masses marked as lobular carcinoma in situ were removed from the dataset. All cysts in the dataset were classified as solitary cysts by experienced radiologists. All images with multiple cysts were removed from the dataset, even though they were recalled in screening. The dataset contained three cases with a breast implant, which were removed from the set as well.

In both datasets, we extracted patches of 260×260 at 200 micron, or $5.2cm$ per patch (unless mentioned otherwise in the experiments section). The centre of each patch was taken as the mean x and y values of a contour drawn by research assistants under the supervision of experienced radiologists.

5.3 Methods

5.3.1 Pretraining Deep Convolutional Neural Networks

Many deep learning applications enjoy gains in terms of improved accuracy or lower training times by employing a network trained on large annotated datasets such as ImageNet [220] and subsequently applying it off-the shelf or finetuning it for the particular task [198, 195], in particular when training data for the actual task is scarce. A similar strategy was also used for several medical problems [8, 46, 121, 234, 236].

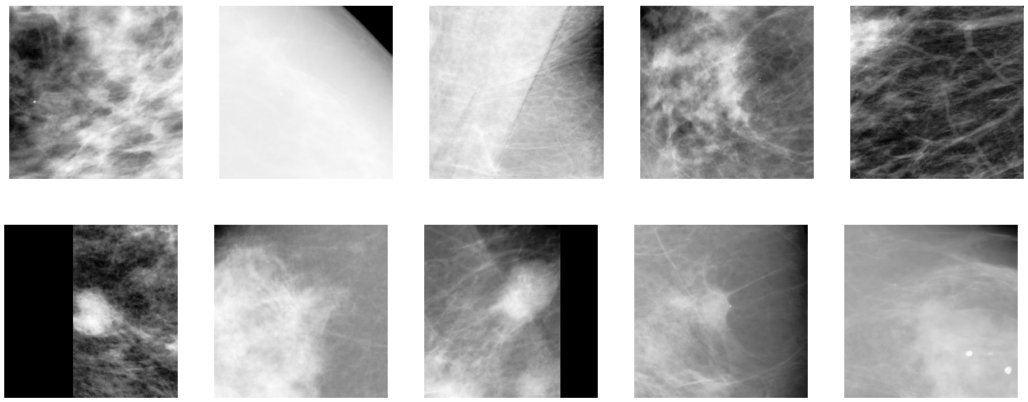


Figure 5.2: Examples of false positive (top row) and true positive (bottom row) lesion candidates. Since the set of solitary cysts and malignant masses is relatively small compared to the amount of data deep neural networks typically need, we first train a network on a big set of mass candidates taken from a large database of screening mammograms.

An arguable downside of using a network trained on natural images is that the input data is significantly different. Natural images often have three channels, whereas medical images can have any format, ranging from grey scale to volumetric, meaning channels need to be copied thereby wasting network parameters. Also, learned invariance properties may not transfer well or may not have been learned from the source dataset. For instance, objects classes in natural images are typically invariant to linear intensity changes, whereas many medical imaging problems are not. Conversely, natural objects are generally not rotation invariant and tumours in mammography are to some extent. Learning features on a large dataset of a more related task could therefore make transfer easier.

Although annotated data of abnormalities in many medical imaging problems is often difficult to come by, normal data can be far more ubiquitous. A possible approach to initialise a network would be to simply train a set of stacked Autoencoders on randomly selected normal samples and subsequently finetune the network for the real classification task. The actual information in normal samples, however, may be low and many weights in the network will be dedicated to represent structure without any discriminative power for the actual problem. Instead, we propose to train fully supervised on a large set of mass candidates: patches found by a candidate detector to resemble masses, extract feature representations from the network and learn a new classifier for diagnosis.

To this end, we use a large dataset of unprocessed screening mammograms. All images are log-transformed and inverted and the breast and pectoral muscle area are segmented. We subsequently extract five texture features, two looking for the centre of a focal lesion, two designed to spot spiculae, characteristic of malignant soft tissue lesions and a last feature capturing the optimal response in scale-space [142]. An ensemble of five Multi Layered Perceptrons (MLPs) is then trained on this feature set. Normal pixels for the training set are sampled randomly from normal parts in the segmented breast area, positive datapoints are sampled densely from a circle of constant size, inside each annotated malignant mass. Pixels in all images are classified using the learned ensemble to form a likelihood image on which we perform non-maximum suppression to generate a set of candidate locations for each image. Centred at each location we extract patches, which are then used to train the deep CNN. Figure 5.2 shows examples of true and false positives the network is trained with.

Rather than retraining the full network, we treat the CNN as a feature extractor and simply extract latent representations from the network by feeding cyst and mass patches and train a shallow non-linear classifier on this feature set. This has as main advantage that the experiments can be run in cross-validation in reasonable time and that the hyperparameters of contemporary ('shallow') classifiers such as Gradient Boosted Trees (GBT) are easier and faster to tune. The latent representation can be accessed from any part in the network, but the dimensionality increases closer to the input patch, imposing computational and memory problems. Instead, we focus on the last three layers. Figure 6.4 shows an illustration of the network and the layers at which we extract features.

An arguable shortcoming of the deep neural network models is that prior knowledge, such as knowledge about invari-

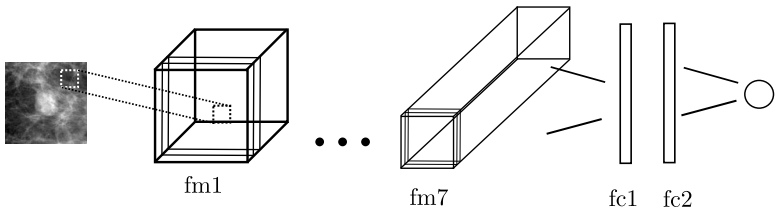


Figure 5.3: Illustration of the deep Convolutional Neural Network (CNN) employed. The network is trained on a large dataset of potential mass regions from screening mammograms (see figure 5.2). Solitary cyst and malignant mass patches from a diagnostic dataset are subsequently presented to the network and features are extracted from the hidden layers and used for the diagnosis task. The network employed is similar to the one in [153, 149]. For brevity only the first feature maps and fully connected layers used are shown. The exact architecture is described in section 5.4. Abbreviations fc1, fc2 and fm7 refer to the fully connected layers, feature maps and their index.

ances to nuisance factors in the classification problem, other than translation, are difficult to encode into the architecture. Rather, CNNs are typically fed with data that is transformed in such a way that it reflects as many possible variations that we expect to see in the image in the hope that the network learns an invariant representation with respect to these nuisance factors. The process of adding training samples this way is known as *data augmentation*. For natural images, simple scaling, translation and colour transformations are typically applied, but for medical data, different sources of variation are present. One of such sources is variation in the amount of tissue surrounding a tumour. All other things being equal, the same tumour will look different on the mammogram if fibroglandular is added or removed. To simulate different forms of occluding tissue, we propose to use *tissue augmentation*, where parenchymal patterns from different images are added to the patch in a physically plausible fashion.

5.3.2 Tissue Augmentations

To perform tissue augmentation, we manually selected 200 patches from normal regions in mammograms of normal breasts recorded with the same detector. In doing so, we made sure the patch was sufficiently far from the breast boundary and that the full patch contained tissue (i.e., no pectoral muscle or boundaries of the breast). To perform realistic augmentations, we make use of the following model of pixel value $I(\mathbf{x})$ at 2D location $\mathbf{x}$:

$$I(\mathbf{x}) = e^{-h(\mathbf{x})\mu(\mathbf{x})} \tag{5.1}$$

with h the height of the tissue and μ the linear attenuation coefficient. By taking the log and adding the augmentation patch also in log space, super scripted a, to source image s we get

$$\log[I(\mathbf{x})] = -h^s(\mathbf{x})\mu^s(\mathbf{x}) + -h^a(\mathbf{x})\mu^a(\mathbf{x})$$

Therefore, physically plausible augmentations entail a simple addition. To prevent unrealistically thick patches, we introduce a blending factor α that governs the amount of tissue added to the source patch:

$$\log[I_t] = \log[I_s] + \alpha \log[I_a] \tag{5.2}$$

Figure 5.4 shows several examples of normal patches (top row) which are added to the first image in the bottom row to generate augmented patches shown in the rest of the bottom row. Each patch in the training set was blended with eight randomly selected patches from the pool of normals. We have chosen to apply the tissue augmentations in the fine tuning stage and not to the pretrained network, such that the effect of the α parameter can be evaluated more easily.

To the best of our knowledge, no methods have been published that tackle this specific problem in mammography. We therefore compare the system to our own previous work, as described in [150] and the chapter 3 of this thesis.

5.4 Experiments

5.4.1 Deep CNN Learning Settings

The CNN was implemented in Theano [16]. We used VGG-like network architectures [246] that was similar to the one employed in Kooi et al. [153, 149] with 2 additional convolutional layers. The model employs 3×3 kernels in

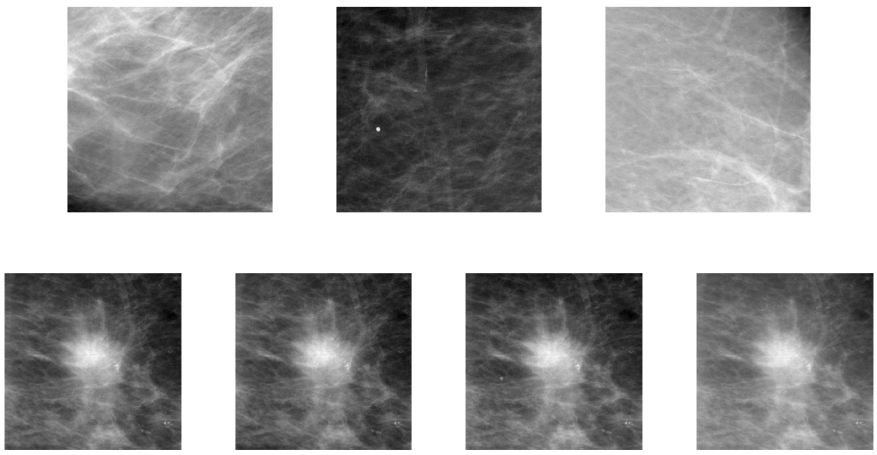


Figure 5.4: Deep CNNs are largely problem agnostic methods. To make them learn the right invariance properties, data augmentation is typically performed, where patches are transformed in such a way that they represent all possible variation in the data. We propose a data augmentation method dubbed *tissue augmentation*, were we randomly select normal tissue from normal areas in the breast and superimpose these over mass and cyst patches to simulate different amounts of parynchemal tissue surrounding the lesions.
(Top row) Normal patches that are superimposed on the leftmost patch in the bottom row generating patch 2-4 in the bottom row. We used a blending factor $\alpha = 0.4$

7 convolutional layers with $\{16, 16, 32, 32, 64,\ 128, 128\}$ feature maps, pooling on $\{2, 1, 2, 1, 2, 1, 2\}$ and ReLU units in all layers, except for the classification layer. Two fully connected layers with 300 units each were added. We have experimented with deeper networks and more feature maps also but did not see an improvement in performance on the validation set. We employed a binary cross-entropy loss which was optimised using ADAM [144] and a learning rate of 0.00005. Dropout [252] was applied to all fully connected layers with $p = 0.5$. We added an L2 norm with a weight of 0.00005 to all layers. Weights were initialised using the MSRA weight filler [113]. Early stopping was used using the AUC on the validation set.

Since the dataset is heavily imbalanced, we generated two separate sets: one with all normal samples and one with all malignant masses. We iterate through the set of normals, reading chunk by chunk from disk and hold the set of positives in host RAM. For each minibatch of normals, we randomly sample an equal amount malignant masses from the set in host RAM to maintain a 50/50 class ratio. The model trained for five days on a Titan X, 12 GB GPU. For our pretrained network, we augmented each patch 16 times with scaling, translation and all 8 reflective symmetry permutations.

Patches were scaled between 0 and 1, when employing tissue augmentations the scaling factors were multiplied by $1 + \alpha$. Since some candidates occur at the border of the breast, we padded each mammogram with zeros.

5.4.2 Top Layer Learning Settings

For both the manual features and the features extracted from the CNN, we employed a Gradient Boosting Tree (GBT) classifier [88] with a binomial deviance loss function:

$$\mathcal{L}(y, h(\mathbf{x})) = \log\left[1 + \exp(-2yh(\mathbf{x}))\right] \tag{5.3}$$

a tighter upper bound on the zero-one loss and less susceptible to outliers than the exponential or squared loss. We employed nested cross-validation with 8 inner folds on the maximum depth and shrinkage and 16 outer folds for testing. We used the square root heuristic for the maximum number of features. Data was split on a patient level to prevent any bias in both inner and outer fold. In all settings, 100 estimators were used. A weight was used inversely proportional to the class ratio to account for imbalance.

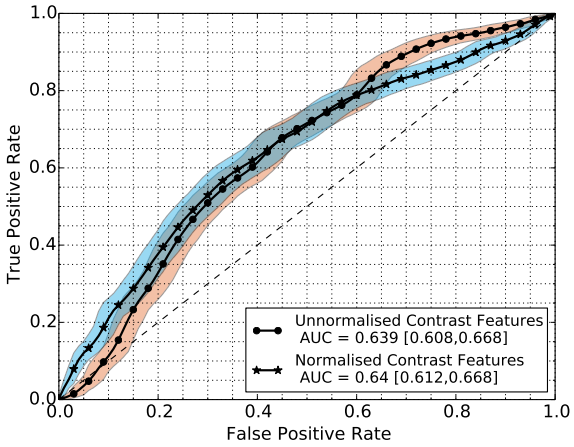


Figure 5.5: ROC results plus a 95% confidence interval of the different sets of contrast features using only a single view. Normalised features refers to the features in the previous chapter, unnormalised features as defined in equations (5.4) and (5.5)

During screening, two images are typically recorded of each breast: a top-down view (CC) and side ways view (MLO), which gives the radiologists additional information. To harness this information, we also experimented with combining the classifiers posterior of the two different views. We evaluated two simple rules: mean and max. If only one view is present or the lesion is only annotated in one of the two views, only this view was used.

5.4.3 Results

As was done in our previous work [150], we first compare the tissue normalised descriptors to two similar naive contrast features: A simple estimate of the mean

$$\mathbb{E}[F_z + \epsilon] - \mathbb{E}[\epsilon] \tag{5.4}$$

and variance:

$$Var[F_z + \epsilon] - Var[\epsilon] \tag{5.5}$$

Figure 5.5 shows the ROCs of obtained with the different sets of contrast features. Figure 5.6 shows results obtained when combining vies using the max rule. 95% Confidence intervals and p-values were computed using bootstrapping [69] with 2000 bootstraps in all settings. We found no significant difference ($p = 0.4456$, mean difference 0.0017 with 95% confidence interval $[-0.0306, 0.0325]$) between the unnormalised and normalised features when using only a single view, but did find a significant difference when combining them using the max rule ($p = 0.0286$).

Effect of the Depth We subsequently investigated the optimal level at which features can be extracted. When extracting features deeper in the network, the size of the feature vector increases, thereby adding some computational burden, so if two depths perform equal in terms of classification performance, the smaller representation is still preferred. Additionally, it gives insight into how 'transferable' the latent representation is, i.e., if feature maps closer to the input give optimal performance it would suggest the learned features are suitable for both tasks.

Table 5.2 shows the AUC values and a 95% confidence interval, obtained after retraining the network with features extracted from the final fully connected layer (fc2), the first fully connected layer (fc1) and the last feature maps (fm7). Single and combined refers to using only a single view and combining two views, as described earlier. We have experimented with feature maps extracted deeper in the network but did not see a clear increase in performance. We found a significant difference between the performance of last feature maps (fm7) and the normalised features, both using single view and the combination ($p = 0.01$).

We did not find a clear difference between using the mean and max rule on this feature set and therefore only the max rule is shown in the table. Although there is an improvement when going deeper and combining views, we also did not find this to be significant in any combination. Figure 5.7 shows the ROC curves comparing the single and combined

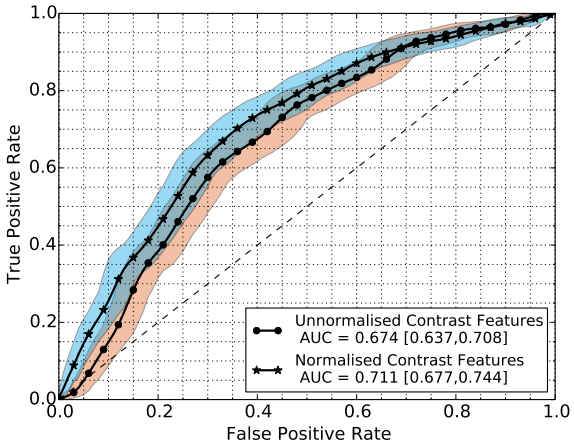


Figure 5.6: ROC results plus a 95% confidence interval of the different sets of contrast features when combining CC and MLO views using the max rule.

Table 5.2: AUC values and 95% confidence interval obtained after extracting features from different layers in the network using patch resolution of 200 micron and retraining a classifier for the task at hand. Abbreviations fc and fm indicate the index of the fully connected layer and feature map, respectively. See figure 6.4 for an illustration. Image based refers to the results obtained after classifying every lesion individually and exam based to the results obtained after combining the output of the classifier for the CC and MLO of the exam.

Layer	Nr of features	Image based	Exam based
fc2	300	0.741 [0.717, 0.764]	0.748 [0.715, 0.780]
fc1	300	0.744 [0.720, 0.767]	0.752 [0.719, 0.783]
fm7	21632	0.767 [0.747, 0.79]	0.773 [0.742, 0.802]

classification with features extracted from the final feature maps.

Effect of Resolution In the next setting, we extracted the cyst and mass patches at 100 micron rather than 200, but kept using the same network trained at a resolution of 200. We used the fully convolutional approach described by Long et al. [178] to also extract the representation from the fully connected layers. This test can give insight into two things: (1) is there any possible additional information when going for a higher resolution and (2) what kind of features are actually learned by the network. Table 5.3 show AUC values and 95% confidence intervals obtained after extracting features from different layers in the network, this time retrained on patches of 100 micron. Results show the higher resolution gives very marginal improvements, but they are not found to be statistically significant ($p \gg 0.05$) and considering the far higher computational needs, we continued the rest of the experiments with the 200 micron patches.

Effect of Mixing We subsequently investigated the effect of the mixing parameter α. Table 5.4 shows AUC values plus again a 95% confidence interval obtained after bootstrapping. Figure 5.8(a) shows the ROC curve obtained when applying the best α (last row in table 5.4). Although we see an improvement using several values of α, we did not find these to be significant when comparing them to the performance without ($p = 0.9$ for the best value, 0.0470 95% confidence interval $[0.0705, 0.0240]$).

Table 5.3: AUC values and 95% confidence interval obtained after extracting features from different layers in the network using patch resolution of 100 micron and retraining a classifier for the task at hand. Abbreviations fc and fm indicate the index of the fully connected layer and feature map, respectively. See figure 6.4 for an illustration.

Layer	Nr of features	Image based	Exam based
fc2	86700	0.762 [0.74, 0.784]	0.775 [0.74, 0.805]
fc1	86700	0.765 [0.742, 0.787]	0.775 [0.744, 0.805]
fm7	107648	0.769 [0.746, 0.791]	0.777 [0.745, 0.806]

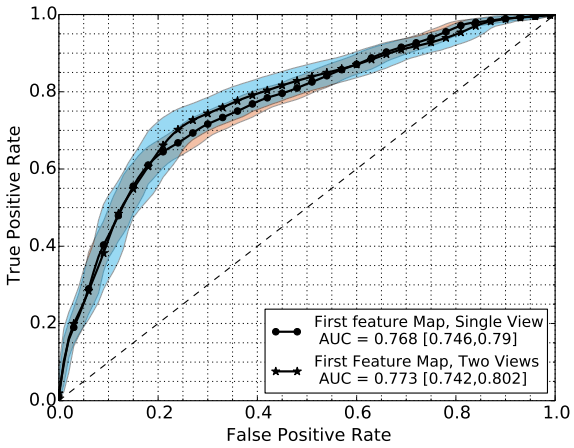


Figure 5.7: ROC curves plus 95% confidence interval obtained after using the features extracted from the last feature map of the CNN. (final row in table 5.2)

Table 5.4: Performance of the system after tissue augmentation, varying the blending factor α (see section 5.3.2). The second and third column show the mean AUC plus a 95% confidence interval obtained after bootstrapping.

α	**Image based**	**Exam based**
0.8	0.762 [0.739, 0.785]	0.779 [0.748, 0.810]
0.72	0.764 [0.741, 0.786]	0.785 [0.753, 0.816]
0.63	0.761 [0.738, 0.784]	0.784 [0.752, 0.814]
0.55	0.762 [0.739, 0.785]	0.783 [0.750, 0.813]
0.47	0.757 [0.734, 0.78]	0.777 [0.745, 0.807]
0.38	0.736 [0.712, 0.76]	0.755 [0.722, 0.786]
0.3	0.73 [0.706, 0.753]	0.749 [0.716, 0.781]
0.22	0.733 [0.709, 0.757]	0.75 [0.716, 0.782]
0.13	0.757 [0.735, 0.780]	0.777 [0.745, 0.807]
0.05	0.775 [0.753, 0.797]	0.795 [0.764, 0.824]

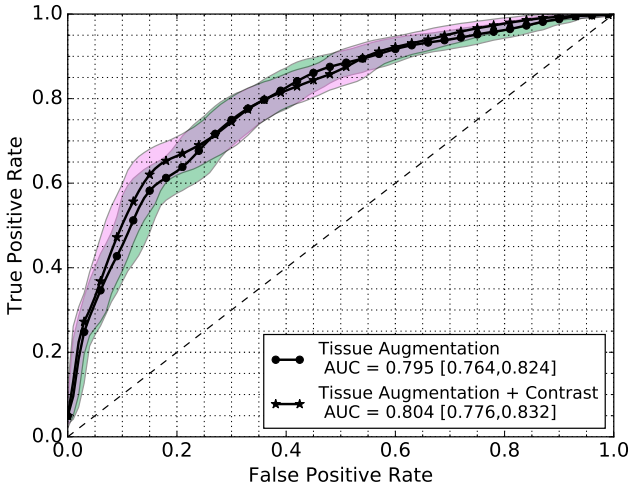


Figure 5.8: ROC curves obtained after using the features from the last feature map and tissue augmentation with $\alpha = 0.05$ and the same configuration after combining it with the manually derived contrast features. We see small improvements by combining the methods but these were not found to be significant.

Table 5.5: Performance of the system for different lesion sizes. The second column indicates the amount of samples, with the amount of malignant masses and benign solitary cysts between brackets. The third column show the performance for different lesion sizes, using the $\alpha = 0.05$, the fourth column the same system but with added contrast features.

Diameter range	**Nr of samples**	**AUC**	**AUC with contrast**
$0 - 10mm$	196 (43/25)	0.753 [0.624, 0.87]	0.776 [0.658, 0.878]
$10 - 13.5mm$	192 (110/66)	0.826 [0.760, 0.884]	0.784 [0.713, 0.850]
$13.5 - 17mm$	162 (132/87)	0.714 [0.643, 0.781]	0.835 [0.782, 0.885]
$17 - 20mm$	109 (111/65)	0.805 [0.729, 0.875]	0.782 [0.707, 0.851]
$20 - 27mm$	163 (136/89)	0.818 [0.757, 0.873]	0.810 [0.751, 0.866]
$> 20mm$	134 (54/38)	0.866 [0.788, 0.932]	0.826 [0.736, 0.904]

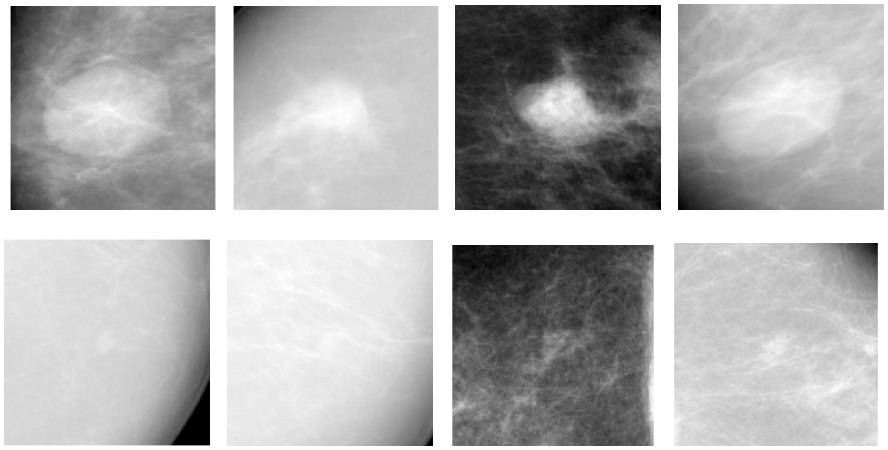


Figure 5.9: (Top row) Cysts that were classified as most like masses by the final system in single view mode.
(Bottom row) Malignant masses that were classified as most like solitary cysts by the final system in single view mode. Very large cysts and very small masses are most difficult for the system. We suspect these are simply underrepresented in the training set.

Effect of Combining with Contrast Features In an attempt to further increase the performance of the system, we combined the normalised contrast features with the features extracted from the pretrained CNN by simply concatenating the feature vectors. For this, we used the value of α found to be optimal in the previous experiment. We obtained an AUC of 0.784 with 95% confidence interval $[0.763, 0.805]$ for the single view setting and an AUC of 0.804 with confidence interval $[0.776, 0.832]$ when combining views. Plots of the ROC curves are provided in figure 5.8.

Effect of size Lastly, similar to Erhard et al. [72], we split the performance of the system over different sizes of lesions. Results for the 200 micron case at fm7 are shown in table 5.5. The AUC for all lesions of size greater than $10mm$ was 0.806 with confidence interval $[0.777, 0.834]$. Figures 5.9 show the cysts and masses that were found to be most difficult by the final system.

5.5 Discussion

From the ROC curves in figures 5.6 we can see the normalisations as described in section 4.3 improve performance when combining views, but not for the single view case. This is contrary to what we found in previous work. Part of this can be explained by the fact that the dataset used in this study is more difficult and in previous work only images were used where two views were present.

It seems the improvement obtained when combining views is less when the method works better on a single view. We believe a big part of this can be explained by the fact that the closer you get to the optimal information a system gets out of a dataset, the more difficult it is to improve upon the results.

From the results in table 5.2, we can see marginal improvements in performance when extracting latent representations deeper in the network. We have tried extracting features closer to the input as well, but did not see an increase in performance. This suggests the tasks are related enough for easy feature transfer and is in contrast to results reported when transferring between natural images and medical images [263], where retraining deeper has been shown to be beneficial for several tasks. A downside of our approach is that the dimensionality of the feature vector increases the deeper the features are extracted. We think retraining the full network could result in a minor performance increase, but due to the difficulty of running this in cross-validation, we have chosen not to do this.

Interestingly, the performance was roughly similar at 100 micron. Since we did not see a very clear increase in performance either, we have chosen to use the downscaled version only. However, the roughly equal performance gives the idea that the features learned are not responding to any specific part but are more general texture descriptors useful for mammography, which we think is an interesting finding. In future work, we plan to explore training a network on mass candidates at 100 micron and subsequently retraining it for this task to see if the performance increases.

From the results in table 5.4 we see the tissue augmentations give a small amount of improvement for large values of α and the most improvement over the baseline for an $\alpha = 0.05$, though none of these improvements were found to be significant. In this study, we have shown results for all values of the mixing coefficient to get some insight into its performance on a test set. When using the system, the parameter should be cross-validated, similar to the parameters of the classifiers. We do believe that this will still give an improvement in performance. Similarly, we found the scaling factors to be an important influence, for which we now choose one setting, but these can equally be cross-validated for better performance.

In recent years, several groups have started working on deriving networks invariant to basic geometric transformations [28, 51, 60] possibly obviating the need for geometric data augmentation. We feel the tissue augmentations proposed are particularly relevant, since it may prove to be difficult to derive networks invariant to this nuisance factor mathematically.

In comparison to a similar study performed with spectral mammography, a specialised modality, Erhard et al. [72] obtained an AUC of 0.88. Their set contained only 62 solid and 52 cystic lesions. From the solid lesions, 15 were benign. Similar to their study, we find the method works slightly better for larger lesions, though very small and very large lesions are most difficult for the system (see figure 5.9). We suspect these are simply underrepresented in the training set and as training data becomes more ubiquitous the performance will increase.

From the curves in figure 5.8, we can see that the performance already comes close to that reported by Erhard et al. [72] and that part of the lesions can be filtered out and do not need to be recalled. We think the proposed deep CNN approach shows great potential and even better results could be obtained. For instance, by looking at table 5.3 we can see a very small increase in performance when using patches extracted at 100 micron. By using a network also pretraining on mass candidates at a resolution of 100 micron and subsequently using the features from this network, we suspect the results to be better and more information can be extracted from the patches. Unfortunately, the increased training time is currently still prohibitive to execute this. Lastly, the context of the mammogram has not been taken into account yet. When radiologists look at an exam, they will most likely not only consider a patch, but look at the image as a whole, which could result in better diagnosis.

5.6 Conclusion

In this chapter, we have presented a Computer Aided Diagnosis (CADx) method to discriminate cysts from solid lesion in mammography using features from a deep Convolutional Neural Network (CNN) trained on a large set of mass candidates, obtaining and AUC of 0.8 on a set of mass candidates recalled from screening. We have compared the CNN based system to our own previous work and only method published for this problem and have shown it outperforms this method. Contrary to related work investigating transfer [236, 263], we do not use pretrained networks on natural images, but use a large dataset from a more related task.

We have shown that by augmenting the patches with randomly sampled tissue from normal images, small improvements in performance can be obtained. The final system works best for lesions larger than 20mm where it obtains an AUC of 0.866. The AUC of the system comes close to the AUC obtained with the recently proposed spectral mammography [72], which is promising. We also believe that with more data and more computational power, the performance can still

be improved significantly.

Chapter 6

Classifying symmetrical differences and temporal change using deep convolutional neural networks

To appear in:
Classifying Symmetrical Differences and Temporal Change for the Detection of Malignant Masses in Mammography Using Deep Neural Networks - *Thijs Kooi and Nico Karssemeijer* - Journal of Medical Imaging, 2017

Abstract

We investigate the addition of symmetry and temporal context information to a deep convolutional neural network (CNN) with the purpose of detecting malignant soft tissue lesions in mammography. We employ a simple linear mapping that takes the location of a mass candidate and maps it to either the contra-lateral or prior mammogram and regions of interest (ROI) are extracted around each location. Two different architectures are subsequently explored: (1) a fusion model employing two datastreams where both ROIs are fed to the network during training and testing and (2) a stage-wise approach where a single ROI CNN is trained on the primary image and subsequently used as feature extractor for both primary and contra-lateral or prior ROIs. A 'shallow' gradient boosted tree (GBT) classifier is then trained on the concatenation of these features and used to classify the joint representation.

The baseline obtained an AUC of 0.87 with confidence interval $[0.853, 0.893]$. For the analysis of symmetrical differences, the first architecture where both primary and contra-lateral patches are presented during training obtained an AUC of 0.895 with confidence interval $[0.877, 0.913]$ and the second architecture where a new classifier is retrained on the concatenation an AUC of 0.88 with confidence interval $[0.859, 0.9]$. We found a significant difference between the first architecture and the baseline at high specificity with $p = 0.02$. When using the same architectures to analyze temporal change we obtained an AUC of 0.884 with confidence interval $[0.865, 0.902]$ for the first architecture and an AUC of 0.879 with confidence interval $[0.858, 0.898]$ in the second setting. Although improvements for temporal analysis were consistent, they were not found to be significant. We feel the results show our proposed method is promising and think performance can greatly be improved when more temporal data becomes available.

6.1 Introduction

During a mammographic exam, images are typically recorded of each breast and absence of a certain structure around the same location in the contra-lateral image will render an area under scrutiny more suspicious. Conversely, the presence of a similar tissue less so. Additionally, due to the annual or biennial organization of screening, there is a temporal dimension and similar principles apply: the amount of tissue is expected to decrease, rather than increase with age and therefore, novel structures that are not visible on previous exams, commonly referred to as *priors*, spark suspicion.

In medical literature, an asymmetry denotes a potentially malignant density that is not characterized as a mass or architectural distortion. Four types are distinguished: (1) a plain *asymmetry* refers to a density lacking convex borders, seen in only one of the two standard mammographic views, (2) a *focal asymmetry* is visible on two views but does not fit the definition of a mass, (3) a *global asymmetry* indicates a substantial difference in total fibroglandular tissue between left and right breast, (4) a *developing asymmetry* refers to a growing asymmetry in comparison to prior mammograms [241, 294]. These types are generally benign, but have been associated with an increased risk [232] and are sometimes the only manifestation of a malignancy. To the best of our knowledge, no relevant work has been done that compares reader performance of malignancies with and without left and right comparisons, but asymmetry is often mentioned by clinicians as an important clue, also to detect malignancies that are classified as a mass. The merit of temporal comparison mammograms on the other hand has been well studied and is generally known to improve specificity without a profound impact on sensitivity for detection [267, 32, 280, 216, 291].

Burnside et al. [32] analyzed a set of diagnostic and screening mammograms and concluded that in the latter case, comparison with previous examinations significantly decreases the recall rate and false positive rate, but does not increase sensitivity. Varela et al. [280] compared the reading performance of six readers and found the performance drops significantly when removing the prior mammogram, in particular in areas of high specificity, relevant for screening. Roelofs et al. [216] also investigated the merit of prior mammograms in both detection and assessment of malignant lesions. Their results show performance was significantly better in the presence of a prior exam, but no more lesions were found. They subsequently postulate priors are predominantly useful for interpretation and less so for initial detection. Yankakis et al. [291] additionally investigate the effect of noticeable change in tissue in mammograms. They generated separate sets of current-prior examination pairs with and without noticeable change and observed that recall rate, sensitivity and cancer detection rate (CDR) are higher when change is noted, but specificity is lower, resulting in a higher false positive rate.

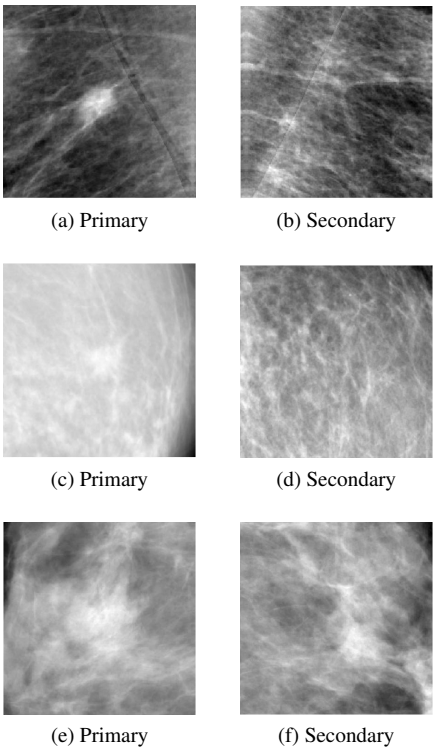
Symmetry is often used as a feature in traditional CAD systems detecting pathologies such as lesions in the brain [174], prostate cancer [171] and abnormalities in the lungs [276]. Most research on mammographic asymmetries involves the classification of a holistic notion of discrepancy rather than the incorporation of this information in a CAD system [80, 35]. Published work on temporal analysis typically relies on the extraction of features from both current and prior exams which are combined into a single observation and fed to a statistical learning algorithm [109, 269]. For detection, an additional registration step is performed [268]. This has been shown to significantly increase performance of the traditional, handcrafted feature based systems.

The vanilla CNN architecture is a generic problem solver for many signal processing tasks but is still limited by the constraint that a single tensor needs to be fed to the front-end layer, if no further adaptations to the network are made. Medical images provide an interesting new data source, warranting adaptation of methods successful in natural images. Several alternative architectures that go beyond the patch level and work with multi-scale [77] or video [137, 194, 245] have been explored for natural scenes. In these settings, multiple datastreams are employed, where each datastream represents, for instance, a different scale in the image or frames at different time points in a video. Similar ideas have been applied to medical data, most notably the 2.5D simplification of volumetric scans [208, 219, 218].

In this chapter we extend previous work [152] and investigate the addition of symmetry and temporal information to a deep CNN with the purpose of detecting malignant soft tissue lesions in mammography. We employ a simple linear mapping that takes the location of a mass candidate and maps it to either the contra-lateral or prior mammogram and regions of interest (ROI) are extracted around each location. We subsequently explore two different architectures

1. A fusion model employing two datastreams where both ROIs are fed to the network during training and testing.

2. A stage-wise approach where a single ROI CNN is trained on the primary image and subsequently used as feature extractor for both primary and contra-lateral or prior ROIs. A 'shallow' gradient boosted tree (GBT) classifier is



(a) Primary (b) Secondary

(c) Primary (d) Secondary

(e) Primary (f) Secondary

Figure 6.1: Examples of symmetry pairs. *Top row:* Very suspicious malignant lesion, regardless of its contra-lateral counterpart.
Middle row: Malignant lesion that is more suspicious in the light of its contra-lateral image.
Bottom row: Normal structure that is less suspicious in the light of its contra-lateral image.

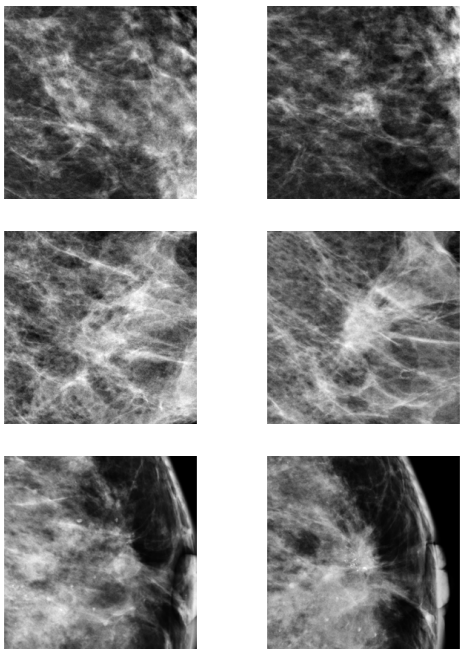


Figure 6.2: Examples of temporal pairs. The right column represents the current and the left column the prior image it is compared with, using the mapping described in section 6.2.2

subsequently trained on the concatenation of these features and used to classify similar concatenations of features in the test set.

Examples of symmetry pairs are show in figure 6.1. Figure 6.2 shows several examples of temporal pairs.

To the best of our knowledge, this is the first CAD and deep learning approach incorporating symmetry as a feature in a CAD system and the first CAD system exploring deep neural networks for symmetry and temporal comparison. Even though the methods are applied to mammography, we feel results may be relevant as well for other medical image analysis tasks, where classification of anomalies that occur unilaterally or develop over time is important, such as lung, prostate and brain images.

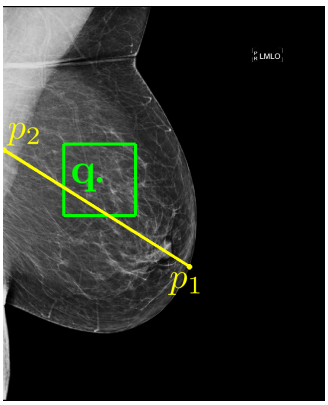
The rest of this chapter is divided into 5 sections. In the following section, we will outline the data pre-processing, candidate detector and linear mapping used. In section 6.3 the deep neural architectures will be described followed by a description of the data and experimental setup in section 6.4. Results will be discussed in section 6.5 and we will end with a conclusion in section 6.6.

6.2 Methods

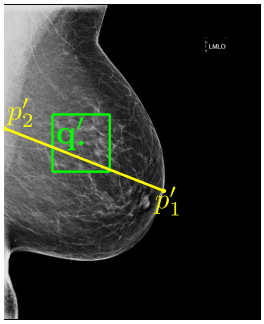
6.2.1 Candidate Detection

We generally follow the candidate detection setup described in Kooi et al. [153]. To get potential locations of lesions and extract candidate patches, we make use of a popular candidate detector for mammographic lesions [142]. It employs five features based on first and second order Gaussian kernels, two designed to spot the center of a focal mass and two looking for spiculation patterns, characteristic of malignant lesions. A final feature indicates the size of optimal response in scale-space. We subsequently apply a random forest [24] classifier to generate a likelihood map on which we perform non-maximum suppression. All optima are treated as candidates and patches of 250×250 pixels, or 5 cm at 200 micron, are extracted around each center location. Since many candidates are too close to the border to extract full patches, we pad the image with zeros.

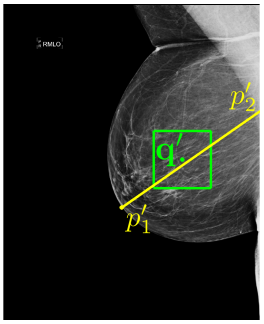
For data augmentation, we follow the scheme described in Kooi et al. [153]. Each patch in the training set containing an annotated malignant lesion is translated 16 times by adding values sampled uniformly from the interval $[-25, 25]$ (0.5



(a) Primary Image



(b) Prior Image



(c) Contra-lateral Image

Figure 6.3: To incorporate symmetry and temporal information, we make use of a simple mapping, based on two coordinates indicated by the end points of the yellow line. (a) A Region Of Interest (ROI) represented by the green box is extracted around a potential malignant lesion location, indicated by the green dot, found by a candidate detector. The location is subsequently matched to either the prior (b) or the contra-lateral image (c). We explore two deep Convolutional Neural Network (CNN) fusion strategies to optimally capture the relation between contra-lateral and prior images.

cm) to the lesion center. Each original positive patch is scaled 16 times by adding values sampled uniformly from the interval $[-30, 30]$ (0.6 cm) to the top left and bottom right of the bounding box. All patches, both positive and negative are rotated using four 90 degree rotations. This results in $(1 + 16 + 16)4 = 132$ patches per positive lesions and 4 per negative. In practice, these operations are computed on the fly during training, to prevent large datasets on disk. After candidates have been generated, locations are mapped to the same point in the contra-lateral image or the prior.

6.2.2 Mapping image locations

Finding corresponding locations between two mammograms is a challenging problem due to two main factors: (1) apart from the nipple and chest wall, which may not always be visible, there are no clear landmarks to accommodate feature based registration and (2) the transformation is highly non-linear. Before the mammogram is recorded the breast is deformed strongly: the viewing area is optimized and dose is minimized by stretching the breast. Additionally, the compression plates may not always touch the breast at the same location causing some movement of tissue within the breast.

A comparative study between several commonly applied registration methods by Van Engeland et al. [274] found a simple linear approach based on the position of the nipple and center of mass alignment outperformed more complex methods such as warping. We propose a similar approach based on two landmarks. To obtain these points, the whole breast area is first segmented using simple thresholding, followed by a linear hough transform to segment the pectoral muscle [138], in the case of an MLO image. The row location of the front of the breast (an approximation of the nipple location) p_1 is subsequently estimated by taking a point on the contour of the breast with the largest distance to the line output by the hough transform. A column point in the pectoral muscle or chest wall p_2 is taken by drawing a straight line

from this point perpendicular to the fit output by the hough transform. The lesion center in the image under evaluation $\mathbf{q} = (q_r, q_c)^T$, where q_r and q_c denote the row and column location, respectively, is subsequently mapped to the estimated lesion center $\mathbf{q}'$ in the contra-lateral or prior image according to:

$$\mathbf{q}' = \mathbf{q} - \mathbf{p} + \mathbf{p}' \tag{6.1}$$

with $\mathbf{p} = (p_1, p_2)^T$ and $\mathbf{p}' = (p'_1, p'_2)^T$ the same points in the contra-lateral or prior mammogram. In other words, we simply clamp the x-distance to the chest wall and the y-distance to the estimated location of the nipple. An example is provided in figure 6.3.

Since most CNN architectures induce a decent amount of translation invariance, the mapping does not need to be very precise. To further mitigate mapping errors, we introduce a form of data augmentation by mapping each location in the image in question to 64 different points in the comparison mammogram by sampling the location from a Gaussian with zero mean and 10 pixel standard deviation.

6.3 Fusion architectures

Partly inspired by the work of Karpathy et al. [137], we propose to add the contra-lateral and (first prior) temporal counterparts of a patch as separate datastreams to a network. In principle, the datastreams can be merged at any point in the network, with simply treating the additional patch as a second channel the extreme case. Neverova et al. [194] postulate the optimal point of fusion pertains to the degree of similarity of the sources, but to the best of our knowledge no empirical of theoretical work exists that investigates this. We evaluate two architectures:

1. A two-stream network where kernels are shared and datastreams are fused at the first fully connected layer. Figure 6.4 provides an illustration of this network.

2. A single patch, single stream network is used as a feature extractor by classifying all samples in the training and test set and extracting the latent representation of each patch from the first fully connected layer $\mathbf{x}^{fc1}$ of the network. This feature representation of the primary and either contra-lateral or prior ROI are concatenated and fed to a 'shallow' GBT classifier to generate a new posterior that captures both symmetry or (first prior) temporal information.

The second approach is far easier to train, since it does not entail re-optimizing hyperparameters of a deep model, which is tedious and time consuming. A downside is that the kernels effectively see less data and are therefore potentially less optimal for the task. Additionally, the second setup is more prone to overfitting. We will elaborate on this in the discussion.

In general, there are a lot less temporal than symmetry samples because they require two rounds of screening and symmetry samples only one. To compare these architectures, we could simply take a subset of the data where each current exam has both a contra-lateral and prior counterpart. Unfortunately, this yields a relatively small number of positive samples and in early experiments, we found the (base) performance to be very marginal and not sufficient to provide a fair comparison. We therefore view missing prior exams simply as missing data. Although missing data has been well studied in the statistics community [4], relatively little has been published with respect to discriminative models.

In the context of recurrent neural networks (RNNs) [105, 107, 169], several imputation methods have been explored [36, 170]. Lipton et al. [170] investigate two imputation strategies: *zero-imputation*, where missing samples are simple set to zero and *forward-filling* that sets the missing value to the value observed before that. Their results show zero imputation with missing data indicators works best, but no significance analysis is performed. In a similar spirit we explore two strategies

1. use a black image when no prior is available. When a woman skipped a screening round, we map the image to the exam four years before the current or add a black image if this is absent.

2. use the image from the exam four years before the current image and use the current when no prior is available.

The first approach carries some additional information, in the sense that the absence of a prior may also increase the likelihood that an exam is positive, since more cancers are typically found in the first round of screening. In the second

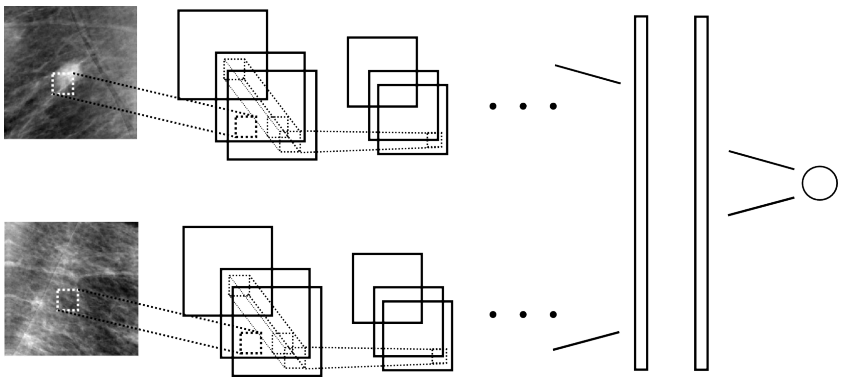


Figure 6.4: To learn differences between left and right breast and temporal change around a candidate location, we use a two-stream Convolutional Neural Network (CNN). The first stream has as input a patch centered at a candidate location, the second stream a patch around the same location in either the contra-lateral image or the prior, using the mapping depicted in figure 6.3. All weights are shared across streams and feature maps are concatenated before the first fully connected layer.

Table 6.1: Overview of the data used for training, validation and testing. Findings refers to the amount of candidates (before data augmentation). Number are separated by '/' where the first number indicates the amount for training, the second the amount for validation and the third the amount for testing.

	Findings	**Cases**
Masses	869/210/470	796/189/386
Normal	200982/54566/74799	3111/1482/1137

setting, it is difficult for the network to distinguish pairs where no change is observed and pairs where simply no prior is available. To add symmetry and temporal information simultaneously, both architectures can trivially be extended with a third stream. However, this requires some additional engineering and we therefore restrict this study to learning two separate models and will propose ways to extend this in the discussion.

6.4 Experiments

6.4.1 Data

Our data was collected from a mammography screening program in The Netherlands (Screening Mid-West) and was recorded with a Hologic Selenia mammography device at an original resolution of 70 micron. All malignant masses were biopsy proven and annotated using contours drawn under the supervision of experienced radiologists. A candidate was considered positive, if the locations was in or within 0.7 cm from an annotated malignant lesion. Before presentation to the human reader, the image is typically processed to optimize contrast and enhance the breast periphery. To prevent information loss, we work on the raw images instead and only apply a log transform which results in a representation in which attenuation and pixel values are linearly related. The images are subsequently scaled to 200 micron using bilinear interpolation.

Our dataset consists of 18366 cases of 18366 women. Each case comprises of one or more exams taken at intervals of two years, unless a women skipped a screening. Each exam again consists of typically four images: a craniocaudal and mediolateral oblique view of each breast. We generated training, validation and test set by splitting on a case level, i.e., samples from the same patient are not scattered across sets. We took 65% for training, 15% for validation and 25% for testing. An overview of the data is provided in table 7.1.

6.4.2 Learning settings and implementation details

The networks were implemented in TensorFlow [1] and generally follow the architecture used in Kooi et al. [153]. We employed VGG-like [246] architectures with 5 convolutional layers with $\{16, 16, 32, 32, 64\}$ kernels of size 3×3 in all

layers. We used 'valid' convolutions using a stride of 1 in all settings. Max pooling of 2×2 was used using a stride of 1 in all but the final convolutional layer. Two fully connected layers of 512 each were added. Weights were initialized using the MSRA weight filler [113], with weight sampled from a truncated normal, all biases were initialized to 0.001. We employed ELU's [49] as transfer functions in all layers. Learning rate, dropout rate and L2 norm coefficient were optimized per architecture. Remaining hyperparameters of all models were optimized on a separate validation set using random search [15].

Since the class ratio is in the order of 1/10000, randomly sampling minibatches will result in very poor performance as the network will just learn to classify all samples as negative. We therefore applied the following scheme. We generated two separate datasets, one for all positive and one for all negative samples. Negative samples are read from disk in chunks and all positive samples are loaded into host RAM. During an epoch, we cycle through all negative samples and in each minibatch take a random selection of an equal amount of positives, which are subsequently fed to GPU where gradients are computed and updated. This way, all negative samples are presented in each epoch and the class balance is maintained. Each configuration trained for roughly 10 days on a TitanX 12 GB GPU.

For the shallow model, we employ Gradient Boosted Trees (GBT) [88] using the excellent XGBoost implementation [39]. We cross-validated the shrinkage and depth using 16 folds. Further parameters were tuned on a fixed validation set using a coordinate descent like scheme. Since the last fully connected layer has size 512, the input to the GBT comprised of 512 features for the single patch setting and a feature vector of 1024 in the symmetry and temporal setting.

6.4.3 Results

Given the results from clinical literature regarding the merit of priors, we focus our results on the classification of candidates and therefore only present ROC curves, rather than FROC curves that are commonly used for detection. To obtain confidence intervals and perform significance testing, we performed bootstrapping [70] using 5000 bootstraps. All curves shown are the mean curve from these bootstrap samples using cubic interpolation. The baseline obtained an AUC of 0.87 with confidence interval $[0.853, 0.893]$.

Figure 6.5 shows the results of the single ROI baseline, and the fusion architectures as described in section 6.3 applied to the symmetry comparison. The first architecture where both patches are presented during training obtained an AUC of 0.895 with confidence interval $[0.877, 0.913]$ and the second architecture where a new classifier is retrained on the concatenation and AUC of 0.88 with confidence interval $[0.859, 0.9]$. We find significant difference at high specificity on the interval $[0, 0.2]$, $p = 0.02$ between the first architecture and the baseline, but no significant difference on the full AUC ($p = 0.14$). For the second architecture we did not find a significant difference between either the baseline or the first architecture.

Figure 6.6 shows the results of the single ROI baseline and the fusion architectures applied to the temporal comparison. We first investigated the difference between the two different strategies to handle missing priors. The approach using the same image obtained an AUC of 0.873 with confidence interval $[0.854, 0.892]$, the approach using the black image for missing priors an AUC of 0.884 with confidence interval $[0.866, 0.902]$. We did not see a significant difference between the strategies $p >> 0.05$, however, the strategy where the black image was used has a higher AUC and we have decided to use this to compare the fusing architectures.
The first architecture where both patches are presented during training obtained an AUC of 0.884 with confidence interval $[0.866, 0.902]$ and the second architecture where a new classifier is retrained on the concatenation an AUC of 0.879 with confidence interval $[0.858, 0.898]$. We did not find a significant difference between any of the architectures $p >> 0.05$, but improvements were found to be consistent during early experiments. Results will be discussed in the following section.

6.5 Discussion

From the curves in figure 6.5 and 6.6 we can see both symmetry and temporal data improve performance, but only see marginal improvements with temporal data. The curves also show the scheme where both ROIs are fed to a single network (architecture (1) in section 6.3) works best. As mentioned in section 6.3, architecture (2) has the advantage that no new networks need to be trained which can take several months to do properly for large datasets. Two disadvantages, however, are that (1) the kernels in the network (parameters up to the first fully connected layer) effectively see less data. In the

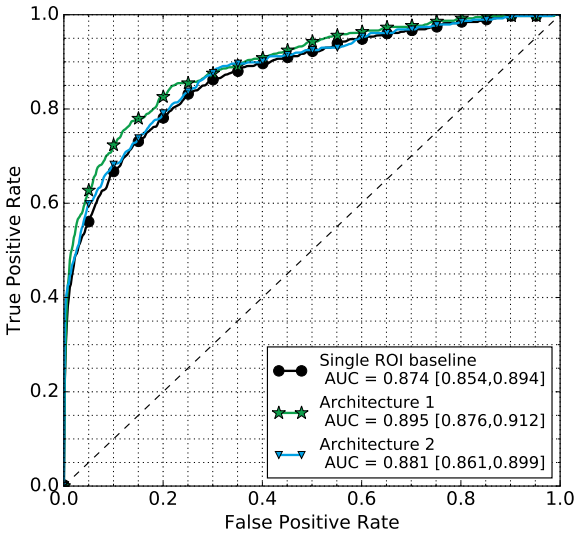


Figure 6.5: ROC curves of the baseline CNN using a single ROI and the two fusing architectures described in section 6.3 when presented with the contra-lateral ROI.

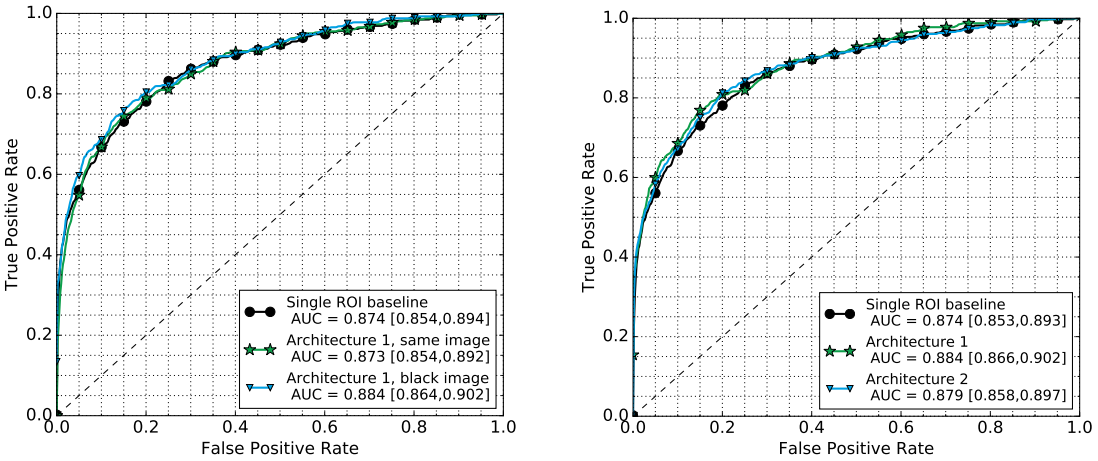


Figure 6.6: (a) ROC curves of the baseline CNN using a single ROI and the two strategies to handle missing prior images both using architecture 1. (b) ROC curves of the baseline CNN using a single ROI and the two fusing architectures described in section 6.3 when presented with the prior ROI and black image strategy.

first architecture, even though the kernels are shared, they are trained on both the primary and either symmetry or prior patch and therefore better adjusted to the task. (2) Overfitting is a much bigger issue: since the features are learned on most of the data the models are trained on, the cross-validation procedure of the GBT often gave a strong underestimate of the optimal regularization coefficients (depth, shrinkage in the case of the GBT), resulting in strong gaps between train and test performance. Optimizing this on a fixed validation set did not result in much better performance. We have tried extracting features from deeper in the network to mitigate this effect but found lower performance.

Since many exams do not have a prior, we explored two strategies to fill in this missing data. In the first setting, we used a black image when no prior image was available and in the second strategy, the same image as the current was used. From the curves in figure 6.6 we can see that in the first setting the prior ROI does add some information and therefore this approach is at least not detrimental to performance. In the second setting, however, we do not see an increase. A possible advantage of the first approach is that it carries some additional information: the number of tumors found in the first screening round is often higher, when using imputation methods mentioned by Lipton et al. [169] this information is effectively lost. As also mentioned in section 6.3, the disadvantage of the second approach is that it is difficult for the network to distinguish between malignant mass-no prior pairs and malignant mass-malignant mass pairs, since no change is typically associated with normal tissue.

In clinical practice, radiologists sometimes look back two studies instead of one, when comparing the current to the prior. Since this requires three screening rounds, this reduces the size of our dataset again, if we want to emulate this and more prior ROIs need an imputed image. Ideally, the neural network architecture should accommodate a varying set of priors. In early experiments, we have explored the use of Recurrent Neural Networks [105, 107, 169], a model designed for temporal data that can be trained and tested on varying input and output sizes. We did not see a clear improvement in performance, but plan to explore this idea more in future work. Since this model can work with varying length inputs, it also provides an elegant way to handle missing prior exams.

In this study, we have trained all networks from scratch. Since the rudimentary features that are useful to detect cancer in one view are expected to be almost as useful when combining views, a better strategy may be to initialize the symmetry or temporal two-stream network with the weights trained on a single ROI. Similarly, since we expect similar features are useful to spot discrepancies between left and right breast as to spot differences between time points, the temporal network could be initialized with the network trained on symmetry patches or the other way around. Due to time constraints this was left to future work, but we suspect an increase in performance.

We have compared two different fusion strategies. As mentioned in section 6.3, the datastreams can in principle be fused at any point in the network, as done by Karpathy et al. [137]. However, there is no guarantee that different architectures perform optimal using the same hyperparameters. For instance, the weight updates of lower layers change if fusion is performed at different points higher in the network. In particular, the learning rate is often found to be important and we feel comparison rings somewhat hollow if no extensive search through the parameter space is done. Since a model typically trains for roughly a week, this is infeasible with our current hardware and we have decided to focus on the two presented models.

Since the focus of this chapter is the presentation of two fusion schemes for adding symmetry and temporal information to a deep CNN, we have presented separate results for each. In practice, when using a CAD system to generate a label for a case, these should be merged into one decision. As mentioned in section 6.3, extending the network with a third datastream is trivial. However, this limits the application to cases where both prior and contra-lateral image are available. In our method, we have added a black image, where priors where not available and a similar approach could be pursued in this setting. Another option would be to train a third classifier on top of the latent representation from separate CNNs or the posterior output by separate CNNs, possibly using a missing data model. Since training deep neural networks and optimizing hyperparameters takes a lot of time, we have left this for future work.

6.6 Conclusion

In this chapter we have presented two deep Convolutional Neural Network (CNN) architectures to add symmetry and temporal information to a Computer Aided Detection (CAD) system for mass candidates in mammography. To the best of our knowledge, this is the first approach exploring deep CNNs for symmetry and temporal classification in a CAD

system. Results show improvement in performance for both symmetry and temporal data. Though in the latter case gain in performance is still marginal, it is promising and we suspect that when more data becomes available, performance will significantly increase. Although the methods are applied to mammography, we think results can be relevant for other CAD problems where symmetrical differences within or between organs are sought, such as lung, brain and prostate images or CAD tasks where temporal change needs to be analyzed, such as lung cancer screening.

Chapter 7

An integrative probabilistic framework for computer aided detection of breast cancer in mammography

Thijs Kooi, Jan-Jurre Mordang, Peter Schulam, Ritse Mann, Nico Karssemeijer and Suchi Saria
In preparation

Abstract

Deep convolutional neural networks (CNNs) have become the model of choice for problems in natural image analysis and the past three years research on medical image analysis is following suit. One difficulty is that medical data is typically too large to feed to a network as a whole and consequently, computer aided detection (CAD) systems have a candidate detection phase to pinpoint salient regions of interest and a classification stage that outputs a score of the particular region. A downside of this approach is that it ignores all context information and any potential interactions between these individual findings. In this chapter, we present a general framework based on a conditional random field (CRF), a model trained on top of the output of different specialized detectors to model interactions between findings. We subsequently propose a method to generate labels for an entire exam, rather than individual findings. The CRF is applied to the detection of breast cancer in mammography and trained on top of state-of-the art mass and calcification detection systems. Results show the model outperforms a state-of-the art system that only works on individual patches. The model is general and can easily be extended with other factors or applied to other CAD problems.

7.1 Introduction

Over the past two years, systems for computer aided detection and diagnosis (CAD) have made a complete shift from using manually defined features to pipelines using deep learning, in particular deep Convolutional Neural Networks (CNN) [172, 108, 74]. These systems made significant improvements upon existing systems [153, 233, 46]. Except for images like color fundus scans [108], or pictures of potential melanoma [74] data is typically too large to feed to a network as a whole. In the case of mammography, a single image can easily contain six million pixels, there are usually four images per exam and multiple exams in a case. Since most of the task relevant information is in-and-around a patch centered at a potential malignancy, feeding the whole case to the CNN effectively decreases the signal to noise ratio, making the problem more difficult to learn. Consequently, many contemporary CAD systems have a detection phase to pinpoint salient regions of interest (ROI) and a classification stage that outputs a score of the particular region [236, 46, 153]. A major downside of this approach is that it ignores all context information and any potential interactions between these individual findings.

Breast cancer has two main manifestations on mammography: malignant soft tissue lesions or masses and calcifications and individual systems are typically developed for each. Microcalcifications are small calcium deposits, originating within the milk ducts and may be associated to ductal carcinoma in-situ or even invasive breast cancer. However, whether they are associated with malignant disease is not easy to determine as many benign changes in the breast present with microcalcifications that strongly resemble the appearance of a malignant disease. Similarly, benign lesions such as cysts, lymph nodes or fibroadenomas can resemble malignant masses, especially when looked at in isolation. On histology, 50 to 80% of masses are found to also contain calcifications and the presence of both findings reinforces the level of suspicion of each individual finding.

On top of the interactions within an image, information outside of the inspected image needs to be taken into account. During a mammographic exam, four images are recorded: a Craniocaudal (CC) or top down view and a Mediolateral Oblique (MLO) or sideways view of each breast. Since screening is typically performed annually or biennially, the structure is augmented with a temporal dimension, all of which contribute to the judgment of readers [267, 32, 280, 216, 291]. To advance the state-of-the art in CAD, systems need to be developed that work on exams and harness all possible information in each image.

Probabilistic Graphical Models (PGMs) [148] facilitate a general framework to reason about different sources of information. PGMs encode a set of independence assumptions between random variables, which are commonly visualized in a graph, hence the name *graphical* model. Directed models or Bayesian networks, found successful application in early expert systems for medical diagnosis such as the MYCIN system. Markov Networks [285] and in particular *Conditional Random Fields* (CRF) [161, 159, 258] are popular models in the computer vision community for reasoning about interactions, such as refining segmentations [91, 239, 92] and object detection [210]. CRFs have recently seen a surge of new interest with several papers showing they can be trained jointly with deep neural networks [40, 231, 165, 299] and have also been applied to medical data [136]. However, they are typically used for segmentation problems in this domain.

The application of PGMs to mammography is not new. In 2009 Burnside et al. [31] already presented a system based on a Tree Augmented Naive Bayes (TAN) [89] model, fed with radiologists findings from mammograms and additional patient covariates such as age, hormone therapy and family history. The network is shown to outperform a set of eight certified radiologists in a retrospective study on the interpretation of nearly 50000 mammographic exams. A downside of this approach is that the TAN is generative and hence requires good approximations of the data distribution. When modeling raw images, this distribution is very high dimensional and therefore complicated to work with. Discriminative models such as deep CNNs circumvent this problem by directly modeling the posterior distribution.

To the best of our knowledge, the only work on combining individual mass findings in a single image was done by Velikova et al. [282, 283], who used a noisy-OR CPD, a type of independence of Causal Influence (ICI) model [114] that approximates the distribution over classes by individually trained models that act through OR gates on the class variable. Though multi-focal mass lesions occur, finding more suspicious mass sites does not make the image more suspicious in general, contrary to pathologies such as diabetic retinopathy, where scores of individual detections can be accumulated [196]. A problem with the noisy-OR model, is that it accumulates evidence, i.e., the more individual findings, the higher the posterior for the full image. A second important issue is that the spatial pattern of individual findings can be of importance: several potential malignant masses close together are more likely to be a group of cysts or lymph nodes. Ad-

ditionally, in the noisy-OR model the set of findings is treated as a bag and the model does not facilitate simple addition of spatial information. Lastly, each finding in an image is associated with one parameter in the model, assuming some order, which in practice may be difficult.

In this chapter, we employ a CRF to integrate different sources of information from a mammographic exam and adopt the model to accommodate posteriors from different classifiers trained on different classes. We subsequently provide a method to generate labels for an exam (a collection of four images) using the output of different detectors and the output of the model. Contrary to the work done by Burnside [31] and Velikova [282, 283], the model provides a general framework by formulating every source of information as a potential function, allowing for easy adjustment if new sources of information, such as additional patient covariates or other imaging modalities, become available and for potentials to be trained discriminatively using popular models such as deep neural networks. Integration of information from multiple detections is a common problem in CAD and we believe the system not only has application in mammography but can also be used in other areas such as lung, retina and prostate CAD.

The rest of this chapter is organized as follows. In the following section, we will introduce the CRF model and the main algorithms we employed. Section 7.3 will describe the application to mammography, followed by the experimental setup and results in section 7.4. We will discuss the results in section 7.5 and will finish with a conclusion in section 7.6.

7.2 Markov Networks and Conditional Random Fields

Markov random fields (MRFs) are a type of probabilistic graphical model and the undirected counterpart of Bayesian networks. Contrary to the latter, the edges between variables are undirected, which is typically more natural for problems in image analysis and additionally allows one to model cyclic dependencies. MRFs, in their general form, model the joint distribution $P(\mathcal{Y})$ of some set of variables $\mathcal{Y} = (Y_1, \ldots, Y_M)$, which typically represent classes of a classification problem. These can be represented as a product of potential functions $\psi(\cdot)$ acting on subsets $\xi_k(\mathcal{Y})$ of variables, where $\xi_k : \mathcal{Y} \mapsto \mathcal{Y}_k$ and $\mathcal{Y}_k \subseteq \mathcal{Y}$:

$$P(\mathcal{Y}|\Theta) = \frac{1}{Z(\Theta)} \prod_{k=1}^{K} \psi_k(\xi_k(\mathcal{Y})) \tag{7.1}$$

where $Z(\Theta) = \sum_{\mathcal{Y}} \tilde{P}(\mathcal{Y}; \Theta)$ a normalization constant known as the partition function. Two types of potentials are often employed for image analysis problems: unary or singleton potentials acting on a single variable and binary or interaction potentials that capture co-occurrence statistics between variables.

Without loss of generality, the potentials can be represented as an exponentiated linear combination of feature functions $\phi_k(\cdot)$ and model parameters $\theta_k^T = (\theta_1, \ldots, \theta_M)$:

$$\psi_k(\xi_k(\mathcal{Y})) = \exp\{\theta^T \phi_k(\xi_k(\mathcal{Y}))\}$$

resulting in a model that can be seen as a structured extension of logistic regression, where instead of a distribution over a single output variable, a joint distribution over a set of variables is learned. The feature functions are often binary mappings but can take any form. Typical functions for segmentations problems in vision are designed to enforce consistency among neighboring pixels, such as $\phi(y_i, y_j) = \mathbb{1}\{y_i = y_j\}$, with $\mathbb{1}\{\cdot\}$ the binary indicator function and the y variables representing pixels in the image.

The CRF model [161, 258], also sometimes referred to as a discriminative random field [159] is a specific type of MRF that assumes every variable Y_k in the model is conditioned on an input $\mathbf{X}_k$. The main advantage in this setting is that parameters in potentials can be trained discriminatively using models like deep CNNs, in which case $\mathbf{X}_k$ is an input patch. This is advantageous if the underlying generative model is complex, but the class posterior relatively simple [159] as in the case of images. Although similar, two computational problems are typically distinguished and are relevant for our application: *inference* and *learning*.

7.2.1 Inference

Inference algorithms are divided into *sampling* based methods that use monte-carlo techniques to approximate the true posterior and *variational* methods that give an exact solution to a tractable surrogate of the true distribution. Both directed

and undirected models can be represented in the form of a *factor graph* [157]: bipartite graphs comprising variable and factor nodes, expediting the generalization of inference algorithms.

A common inference problem is computing marginals: given a joint distribution $P(\mathcal{Y})$ over a set of random variables $\mathcal{Y} = \{Y_m\}_{m=1}^{M}$, compute the distribution $P(Y_m) = \sum_{\mathcal{Y}\neg m} P(\mathcal{Y})$ over individual variable Y_m. These values are needed in our model to eventually generate image based labels. Marginals can simply be computed by summing out all other variables in the distribution. However, the time complexity of this operation is exponential in the amount of variables in the graph and therefore often not possible in practice for all but the smallest models.

Belief propagation [203, 292] is a type of variational inference introduced to efficiently compute marginals [292] and reduces the complexity of the computation from exponential to linear in the amount of variables in the graph. It is phrased as a recursive algorithm that sends messages between nodes in the graph about instantiations y_m of a variable Y_m. In the case of a factor graph, two type of operations are performed: (1) a variable m to a factor k message:

$$\mu_{m\to k}(y_m) = \prod_{k'\in N(m)\neg k} \mu_{k'\to m}(y_m) \tag{7.2}$$

where $N(m)\neg k$ generates the set of all factors containing variable m, excluding k
and (2) factor to variable message:

$$\mu_{k\to m}(y_m) = \sum_{y\in\xi_k(\mathcal{Y})\neg Y_m} \psi_k(\xi_k(\mathcal{Y})) \prod_{k'\in N(k)} \mu_{k'\to m}(y_m) \tag{7.3}$$

with again $N(k)$ a neighborhood generating function, this time returning all variables in the neighborhood. This algorithm will output refined scores for each variable in the model, that take into account any co-occurrence relations and all factors in the model.

7.2.2 Learning

Maximum Likelihood Estimation (MLE) is the most commonly used technique to train PGMs. In the fully observed case, the log-likelihood of parameters Θ conditioned on a dataset $\mathcal{D} = \{\mathbf{X}_n, \mathcal{Y}_n\}_{n=1}^{N}$ under a CRF is given by:

$$\log[\mathcal{L}(\Theta;\mathcal{D})] = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} \psi_k(\xi_k(\mathcal{Y}_n)) - \log\left[Z(\Theta;\mathbf{X}_n)\right] \tag{7.4}$$

where samples are assumed to be iid. Taking partial derivatives with respect to parameters in the model results in the difference between what is referred to as the *clamped* and *contrastive* term:

$$\overbrace{\frac{1}{N}\sum_{n=1}^{N}\phi_k(\xi_k(\mathcal{Y}_n);\theta_k)}^{\text{clamped term}} - \overbrace{\sum_{\mathcal{Y}} p(\mathcal{Y}|\mathbf{X};\Theta)\phi_k(\xi_k(\mathcal{Y}_n);\theta_k)}^{\text{contrastive term}} \tag{7.5}$$

Since $\frac{1}{N}\sum_{n=1}^{N}\phi_k(\xi_k(\mathcal{Y}_n)) = \mathbb{E}_{\mathcal{D}}[\phi_k(\xi_k(\mathcal{Y}_n))]$ the expectation of the feature in the data and $\sum_{\mathcal{Y}} p(\mathcal{Y}|\mathbf{X};\Theta)\phi_k(\xi_k(\mathcal{Y}_n)) = \mathbb{E}_{\Theta}[\phi_k(\xi_k(\mathcal{Y}_n))]$ the expectation of the model, this process is also referred to as moment matching. The CRFs loss function is convex, but has no closed form solution and hence iterative methods, in particular variations on Gradient Descent are applied to get the optimal set of parameters.

The contrastive term in equation (7.5) is exponential in the number of variables K in the graph and due to the dependence on the input in the CRFs formulation, needs to be performed for every training step, rendering learning slow or intractable for large graphs with many edges. Several approximate learning methods [200] have been proposed.

7.2.3 Approximate learning

Popular approximate learning methods include Pseudo-Likelihood (PL) [18], Contrastive Divergence (CD) [118, 34, 287] and piecewise training [259]. Pseudo-likelihood reduced the complexity to polynomial by assuming that all variables

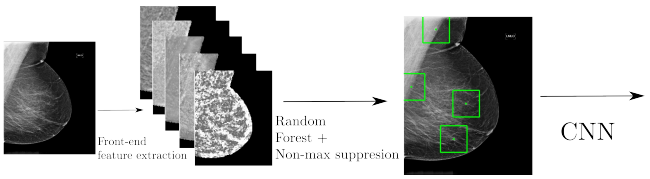


Figure 7.1: We propose to train a conditional random field (CRF) on top of the output of several mass and calcification detection systems. The pipeline used for all mass potentials employs a candidate detector using five Gaussian derivative based features, which are fed to a random forest (RF). The RF is then used to classify pixels and generate a likelihood map of likely mass lesions. A non-maximum suppression is employed to generate local optima, around which patches are extracted that are fed to several deep convolutional neural networks (CNN).

are observed during training. The likelihood is estimated by conditioning all variables on its observed neighbors and subsequently taking an average:

$$P(\mathcal{Y}|\mathbf{X};\Theta) = \prod_{n=1}^{M} P_{PL}(Y_k|\mathcal{Y}_{\neg k},\mathbf{X};\Theta) \tag{7.6}$$

Since the normalization constant now depends on one variable only, the complexity reduces from exponential to linear in the amount of variables in the graph.

7.3 Application to mammography

As mentioned in the introduction, there are two important markers for breast cancer in mammography: mass-like lesions and calcifications and different systems are currently developed for each. However, the co-occurrence of these finding should reinforce the suspiciousness of each individual finding. Additionally, there is symmetry and temporal information that can be used as an additional cue for the class of mass findings [151]. To integrate these source of information we work with two different systems a *mass detection* and *calcification detection* system, each buttressed by a separate pipeline. For the mass detection system we train four different singleton potentials and add one singleton potential from the calcification pipeline to the model. An interaction potential captures the co-occurrence relations between mass findings and mass-calcification pairs.

7.3.1 Singleton Mass Potentials

Similar to many CAD pipelines, we employ a two-stage system with a candidate detection and classification step. All images are processed by first segmenting the breast area and correcting for fall off at the edge of the breast using the peripheral enhancement method proposed by [250]. We subsequently follow the pipeline described in Kooi et al. [153] which we briefly summarize here. A candidate detector [142] is employed that makes use of five features based on first and second order Gaussian kernels, two designed to spot the center of a focal mass and two looking for spiculation patterns, characteristic of malignant lesions. A final feature indicates the size of optimal response in scale-space. A random forest [24] is then trained on this feature set and used to classify the pixels to generate a likelihood image. We perform a non-maximum suppression on these images to generate local optima, which are used as centers of mass candidate locations. This generates a set of mass variables $\mathcal{Y}_{MASS}$, with every $Y \in \{0, 1\}$, i.e., normal or malignant. Cardinality $|\mathcal{Y}|$ is typically in the order of 10-15. The region based pipeline is illustrated in figure 7.1.

For every mass candidate location, we train *four* separate unary potentials: a single patch CNN, a two-stream CNN capturing symmetry, a two-stream CNN capturing temporal information and a location potential:

1. **Single stream mass potential** For the main singleton potential, we extract patches $\mathbf{X}_P$ around each candidate location, following the approach described in Kooi et al. [153]. This generates a single view potential function $\psi_{SV}(Y) = \log\left[P(Y|\mathbf{X}_P;\Theta_{SV})\right]$.

2. **Multi-stream symmetry potential** When radiologists read mammograms, they often compare the left and right breast for potential differences. To capture this in our model, we follow the approach described in [151]. Each candidate is mapped to a location in the contralateral image and two patches $\mathbf{X}_P$ and $\mathbf{X}_S$ are extracted, which are fed as separate streams to a convolutional neural network. The network generates a posterior $P(Y|\mathbf{X}_P,\mathbf{X}_S;\Theta_{SYMM})$

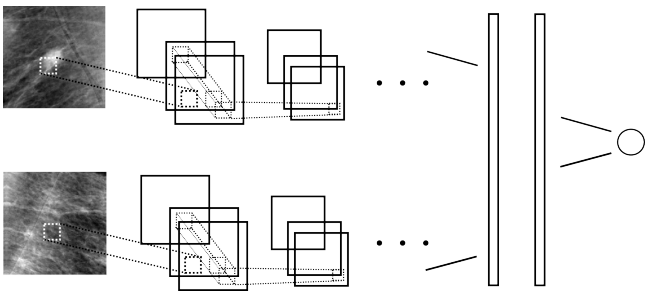


Figure 7.2: Differences between the contralateral and previous mammograms can indicate the presence of a tumor. The model this, we make use of two multi-stream deep CNNs, as presented in Kooi et al. [151]. Patches are extracted around local optima generated by the pipeline described in figure 7.1. These locations are mapped to the contralateral and prior mammogram and fed as separate streams to a two-stream symmetry and two-stream temporal network, trained to classify the top patch. This generates a symmetry and temporal potential, which are used in the CRF.

and corresponding potential $\psi_{SYMM}(Y) = \log\left[P(Y|\mathbf{X}_P, \mathbf{X}_S; \Theta_{SYMM})\right]$. An illustration of this network is provided in figure 7.2.

3. **Multi-stream temporal potential** In a similar way, radiologists also compare the current image to prior exams. To model this, we again follow the approach described in Kooi et al. [151]. Similar to the symmetry potential, locations are mapped to the prior exam to generate patch pairs $\{\mathbf{X}_P, \mathbf{X}_T\}$. We subsequently train a deep CNN $P(Y|\mathbf{X}_P, \mathbf{X}_T; \Theta_{TEMP})$ to generate a potential $\psi_{TEMP}(Y) = \log\left[P(Y|\mathbf{X}_P, \mathbf{X}_T; \Theta_{TEMP})\right]$.

4. **Location potential** Cancer is more likely to occur in some areas of the breast and therefore location is a feature that needs to be taken into account. Since we train the CNN on patches, location is not explicitly represented in the model. To capture this in the CRF, we train a classifier $P(Y|\mathbf{x}_l; \Theta_l)$ on a feature vector $\mathbf{x}_l$ of relative location features based on the location of the nipple and the chest wall to generate a potential $\psi_L(Y) = \log\left[p(Y|\mathbf{x}_l; \Theta_l)\right]$. The set of features is described in more detail in Kooi et al. [153].

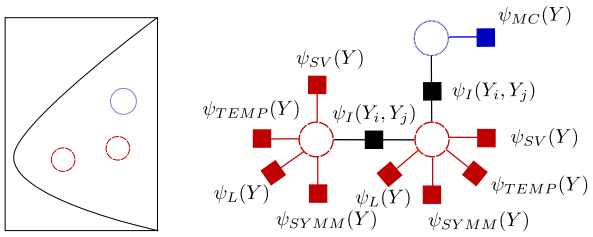
7.3.2 Singleton Calcification Potential

For the calcification pipeline, we generally follow the system described in Mordang et al. [190] which comprises of three main steps. Similar to the soft-tissue pipeline, a pixel classifier is trained on and applied to each mammogram to obtain candidates, individual calcifications in this case. This is followed by a second step where calcifications are segmented with a connected-component analysis. Calcification candidates with a distance less than or equal to 10 mm to another potential site are clustered together to form groups [25]. A set of features based on shape, the likelihood output by the candidate detector, topology, texture and vesselness [281, 25, 191] is subsequently extracted from each cluster and a classifier is trained on this set of features. A final false-positive reduction step is applied to filter out as many benign groups as possible [191]. This gives us a set $\mathcal{Y}_{MC}$ of around three to five calcification findings per image, with an associated potential function $\psi_{MC}(Y)$.

A simple way to combine these systems is to add an additional feature to the final layer of the CNN or set region features from the calcification system, that captures the posterior of other regions in the image. However, reasoning is often an iterative procedure and simply adding features does not capture this. We therefore propose to add an interaction potential to the CRF that captures co-occurrences between findings.

7.3.3 Interaction potential

To summarize: an image contains two different types of variables, mass and calcification candidates, each detected by a different system. In the end, we want to learn a posterior distribution $P(\mathcal{Y}|\mathcal{X}, \Theta)$ over all variables $\mathcal{Y} = (\mathcal{Y}_{MASS}, \mathcal{Y}_{MC})$, where $\mathcal{X}$ denotes the set of all input features or patches. To make modeling easier, we combine the output of the mass and calcification detection into one unary potential $\psi(Y_k)$. To do this, we considered two approaches: (1) conflate the classes into normal and malignant (either mass or calcification) (2) transform the output to a three class problem of normal, malignant mass and malignant calcification classes. In the first setting, learning and inference are more efficient, but this



(a) (b)

Figure 7.3: We present a method to model interactions between findings output by a mass and calcification detector, using a Conditional Random Field (CRF) and to generate labels at exam level that incorporate temporal and symmetry information.
(a) Schematic illustration of a breast with two mass findings (red circles) findings and one calcification (blue circle) finding. (b) The resulting factor graph, where circles indicate variables and squares factors. Each mass finding variable is governed by three factors, a single view CNN, a symmetry CNN and a temporal CNN. The calcification variable is governed by a single potential. The black squares denote an interaction potential that models the co-occurrence of different mass findings and masses and calcifications in a single images.

reduces the amount of information that can be captured. For instance, co-occurrence of mass and calcification candidates can not be captured this way. We therefore opt for the second approach and generate a new singleton potential according to:

$$\psi(Y_k) = \begin{cases} (\psi_{MASS}(Y == 0), \psi_{MASS}(Y == 1), 0) & \text{if } Y \in \mathcal{Y}_{MASS} \\ (\psi_{MC}(Y == 0), 0, \psi_{MC}(Y == 1))) & \text{if } Y \in \mathcal{Y}_{MC} \end{cases} \tag{7.7}$$

This allows us to define an interaction potential between findings i and j in the set $\mathcal{P}$ of all pairs, for which we propose the following:

$$\psi_I(Y_i, Y_j) = \begin{pmatrix} \theta_1 & \theta_2 & \theta_3 \\ \theta_2 & \theta_4 & \theta_5 \\ \theta_3 & \theta_5 & \theta_6 \end{pmatrix} \tag{7.8}$$

i.e., a Potts model, where parameters $\mathbb{1}\{Y_1 = 1, Y_2 = 0\}$ and $\mathbb{1}\{Y_1 = 0, Y_2 = 1\}$, etc. are shared. Since the system is overdetermined, θ_1 is clamped to 0.

7.3.4 Edge construction

An advantage of the CRF is that it provides a way to inject task specific knowledge. In this spirit, we model the following domain knowledge: (1) benign findings such as cysts resemble malignant masses and often occur in pairs. Apart from multi-focal lesions, there are rarely multiple malignant masses in an image. When multiple potential masses are found at a distance larger than one would expect the distance of two foci of a multi focal mass to be apart, they are more likely to be benign and (2) malignant soft tissue lesions and calcifications often occur together, if two suspicious locations are found, they reinforce each other, in particular when they occur in the same segment of the breast. To capture this, we construct edges in the following way:

1. If soft tissue lesion are at a distance larger than 2.5 cm and closer than 6 cm an edge is formed. This way, multiple findings in multi-focal lesions are not connected and findings far away are treated as independent.

2. If soft tissue lesion findings and calcification findings are closer than 3 cm we construct an edge.

An illustration of a mammogram with two mass and one calcification findings and the corresponding factor graphs are depicted in figure 7.3.

7.3.5 Final model

When combined, this gives us the following model:

$$\begin{aligned}&\log\Big[P(\mathcal{Y}|\mathcal{X};\Theta)\Big] = \\ &\sum_{Y\in\mathcal{Y}^{Mass}} \theta_{SV}\psi_{SV}(Y) + \theta_{SYMM}\psi_{SYMM}(Y) + \theta_{TEMP}\psi_{TEMP}(Y) + \\ &\sum_{Y\in\mathcal{Y}^{Mcalc}} \psi_{MC}(Y) + \sum_{(i,j)\in\mathcal{P}} \psi_I(Y_i,Y_j) - \log\Big[Z(\mathcal{X};\Theta)\Big]\end{aligned} \tag{7.9}$$

where $\mathcal{X}$ denotes the set of all inputs (potential mass patches, calcification feature vectors and location feature vectors) and θ_{SV}, θ_{SYMM} and θ_{TEMP} mixing coefficients between singleton potentials.

7.3.6 Aggregating labels

To fully automate the screening process, we need a system that can generate a label for every exam and not simply a vector of labels for all findings output by all individual systems in every image. As a ground truth, an exam is positive if any of the images contains a malignant mass or calcification. Logically this translates to either the OR or MAX rule, i.e., $Y_{case} = \{Y_1 \vee Y_2 \vee \ldots \vee Y_K\}$ or $Y_{case} = \max\{Y_1, Y_2, \ldots, Y_K\}$, which are equivalent for binary labels. However, when employing the OR rule to continuous outputs, the same problem with the noisy-OR model, stipulated in the introduction occurs: the more candidates are found the higher the score of the exam. The MAX rule, on the other hand, requires calibrated classifiers. If the output of the mass and calcification models are not on the same scale, the exam based score can be dominated by one of the two, resulting in poor performance. To prevent this, we first calibrate the classifiers before feeding the posteriors to the CRF.

A classifier is said to be well-calibrated if the empirical class membership converges to the output of the classifier. Intuitively, this means that from all samples where the model assigns a score of 0.8, 80% of these should belong to the positive class [296]. Unfortunately, this does not hold in general and tree based methods such as random forests are known to be especially poorly calibrated. In the past decades, several calibration methods have been suggested. Platt scaling [206] is a commonly used method that essentially uses a logistic regression to recalibrate a model. Isotonic regression [296] is a (restricted) non-parametric method that, contrary to Platt scaling, does not make assumptions about the shape of the reliability graph except that it monotonically increases. It is therefore more widely applicable, but needs more data to get a good fit.

7.4 Experiments

7.4.1 Data

Our data was provided by a screening program in The Netherlands (screening mid-west). All mammograms are recorded using Hologic Selenia devices at an original resolution of 70 micron. For the CNNs, the images were downscaled to 200 micron. Train, validation and test data were split on a patient level to prevent any bias. All normal training data was extracted from normal cases that had a normal screening as follow-up to confirm normality. If a prior has malignant findings that were annotated in hindsight, this is treated as a positive. A mass finding was considered positive if the center is within 1 cm from an annotated contour. For the calcifications, a cluster is considered positive if two or more individual calcifications lie within the annotated contour. In the test set 2649 exams had a prior and in the training set this was 8360. An overview of the data is provided in table 7.1.

7.4.2 Training Settings

Since the CRF is trained on top of a very large pipeline, we employ stage-wise training. The mixing coefficients between singleton mass potentials are first estimated, followed by the calibration and lastly the interaction potential. Below are details of each individual model

Candidate detector, calcification detection, location potential We made use of a random forest (RF) [24] to generate the mass candidates and the mapping in the calcification and location potential. Trees in the RF were grown using the

Table 7.1: Number of exams used for training, validation and testing in the format. The data was collected from a large screening program in the Netherlands. All tumors are biopsy proven and were annotated under the supervision of an experienced radiologist. The data was split on a patient level so that two images of the same patient are never in both the train and test set. About 76% of exams had a prior, which amounts to 8360 in the train+validation set and 2649 in the test set.

	Training	**Validation**	**Test**
Normal	10372	2593	3004
Total malignant	534	133	312
Malignant masses	436	91	209
Malignant calcifications	230	85	139
Malignant mass + calcification	61	15	36

Gini criterion for splitting and in all situations we used 2000 estimators and the square root heuristic for the maximum number of features. The maximum depth was cross-validated using 8 folds. Data was balanced by drawing bootstrap samples with an equal class ratio. The systems are trained using at most the ten most suspicious lesions per image found by the candidate detector. During testing no such threshold is applied to obtain highest possible sensitivity.

Deep CNNs For all CNNs we employ VGG-like models [246], similar to the ones described in Kooi et al. [153, 149, 151], which were implemented in TensorFlow [1]. For all models, we used 5 convolutional layers with $\{16, 16, 32, 32, 64\}$ kernels of size 3×3 in all layers. We used 'valid' convolutions with a stride of 1 in all settings. Max pooling of 2×2 was used using a stride of 1 in all but the final convolutional layer. Two fully connected layers of 512 each were added. Weights were initialized using the MSRA weight filler [113], with weights sampled from a truncated normal, all biases were initialized to 0.001. We employed ELU's [49] as transfer functions in all layers. Learning rate, dropout rate and L2 norm coefficient were optimized per architecture. All other hyper parameters of all models were optimized on a separate validation set using random search [15].

To account for the large class imbalance, which is typically in the order of 1/10000, we generate two separate datasets: one set of normals and one set of malignant masses. The negative samples are read from disk chunk by chunk and all positive samples are read into host RAM at the start of training. During an epoch, we cycle through all negative samples and in each minibatch take a random selection of an equal amount of positives, which are subsequently fed to GPU where gradients are computed and updated. This way, all negative samples are presented in each epoch and the class balance is maintained. We trained and optimized each configuration for roughly four weeks a Titan X 12 GB GPU.

CRF The CRF was implemented using a combination of Matlab and C++ and used the inference algorithms from libDAI [187]. Using a highly optimized implementation with multi-threading the model took several days to train on an IntelI7. We trained the model with SGD with a learning rate of 0.0001 and a momentum term with weighting factor 0.9 and an L2 norm with weight 0.005. Since the CRF's loss function is convex, we did not tune the learning rate except to prevent oscillations (i.e., taking a learning rate sufficiently small to converge). The L2 norm weight was optimized on the validation set. During test time, we applied loopy BP without damping, since the algorithm typically converged after several iterations. We have experimented with Gibbs sampling and mean field variational inference but did not see large differences in performance. We are currently working on a Python version that will be made open-source as soon as possible.

Calibration methods The parameters of the Platt scaling were computing using L-BFGS and no regularization (since input space is one-dimensional). We cross-validated the bin size for the isotonic regression and used the pair adjacency violator (PAV) algorithm to compute the mapping. To handle class imbalance, we employed a weight inversely proportional to the class ratio when using Platt scaling and subsampled the data when using isotonic regression.

7.4.3 Results

Considering how temporal information is used, we present all results as exam (i.e., a collection of CC/MLO pairs of each breast) based curves. We first investigate the effect of the calibration methods described in section 7.3.6 to set a fair baseline. On the level of individual candidates (the mass detection pipeline outputs roughly 8 and the calcification pipeline roughly 4 per image), the single channel network gave an AUC of 0.89 and calcification detector AUC of 0.901.

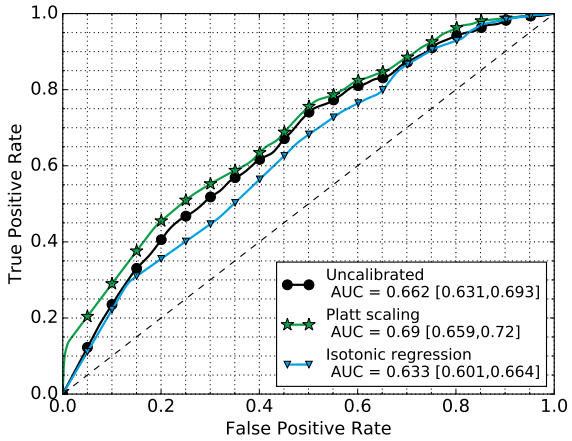


Figure 7.4: To generate labels for an exam, we combine the output of a mass and calcification detection system, for which we propose to take the maximum of all scores from all candidates. In practice, however, the output of a classifier is not scaled in such a way that it can be interpreted as a probability and therefore, when taking the labels as the maximum of two models, one could dominate the final label. To prevent this, we evaluate two scaling methods: Platt scaling and isotonic regression. For a more detailed description see section 7.3.6.

Figure 7.4 shows the exam based ROC curves of three methods described in section 7.3.6. Using the max rule without any calibration, we obtain an exam based AUC of 0.66 with 95% confidence interval $[0.632, 0.692]$, using Platt scaling, the exam based AUC is 0.690 with 95% confidence interval $[0.660, 0.720]$ and using isotonic regression, we obtained an AUC of 0.633 with 95% ci of $[0.601, 0.664]$. Given the superior performance of Platt scaling, all further results use this calibration method. Confidence intervals are computed using bootstrapping with 5000 samples.

We subsequently added each potential to the single channel baseline to finally get the model defined in equation (7.9). The location based potential obtained an AUC of 0.78 on a candidate level, the temporal potential an AUC of 0.90 and the symmetry potential an AUC of 0.91. Figure 7.5 shows the exam based ROCs of the model using only the single channel and microcalcification potential (S), the single channel, location and microcalcification potential (S + L), the single channel, location, temporal and calcification potential (S + L + T), the single channel, location, temporal and symmetry potential (S + L + T + S) and the full model (FULL). The single channel potential combined with the location potential obtained an AUC of 0.706 with 95% confidence interval $[0.674, 0.736]$, when the temporal potential was added, the AUC went up to 0.737 with 95% confidence interval $[0.705, 0.768]$, when adding the symmetry potential, the exam based AUC was 0.752 with 95% confidence interval $[0.721, 0.782]$ and the full model obtained and AUC of 0.755 with 95% confidence interval $[0.723, 0.785]$. Confidence intervals were generated using bootstrapping, with 5000 samples. Examples of inference as a result of the interaction potential are provided in figures 7.6(a) and 7.6(b).

7.5 Discussion

The curves shown in figure 7.4 indicate that using isotonic regression actually decreases performance. During testing, we could see the model overfits, in spite of the cross-validation employed and saw the AUC on the training set increase but the AUC on test set decrease. Additionally, although it is a de-facto non-parametric method, the output of the classifier still needs to be discretized and therefore the bin size is a dial that needs to be tuned. Since Platt scaling is a parametric method, it is less prone to overfitting and we found it easier to train. Handling class imbalance is another challenge. Since employing class weights when using isotonic regression is complicated, we have chosen to subsample the data which, due to the limited amount of malignant samples, gives a very small training set.

From the results provided in figure 7.5, we can see a clear increase from using only the baseline (S) to using the full model (FULL). Although the single channel, temporal and symmetry potential are correlated, since all of the models contain the patch under evaluation, their combination still yields an improvement. In general the performance of the interaction potential is disappointing, since we can see no clear difference in AUC when this is added. Although the potential does what it is designed to do and weights converge to values that make sense intuitively, the added benefit of this in

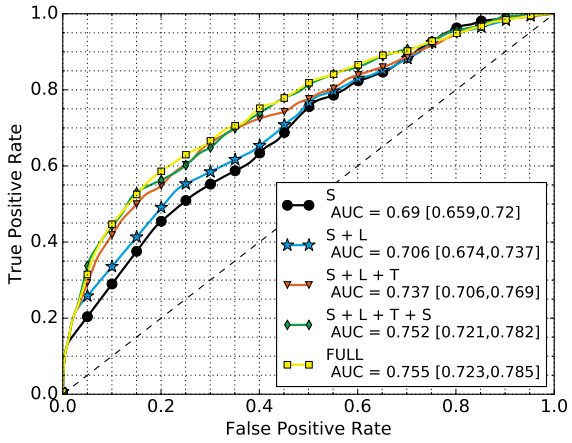
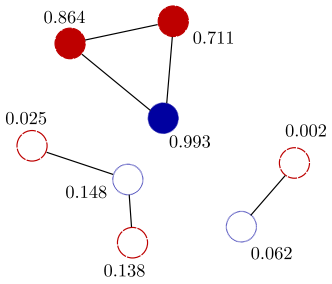
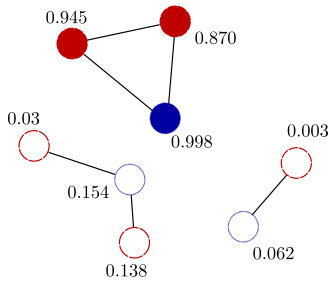


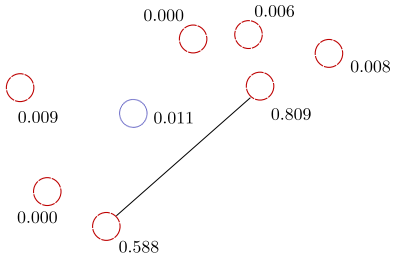
Figure 7.5: Exam based ROCs obtained after adding the potentials described in section 7.3 one by one. S refers using on the potential trained on a single patch, S + L the combination of the single channel and location potential, S + L + T the combination of the single channel, location and temporal information, S + T + L + S the addition of symmetry information and FULL the final model described in (7.9).



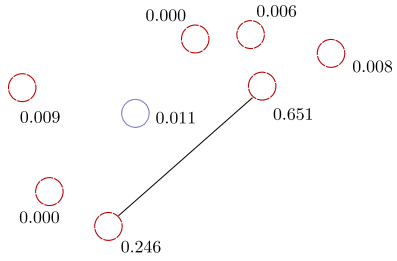
(a) Before inference with the CRF



(b) After inference with the CRF



(c) Before inference with the CRF



(d) After inference with the CRF

Figure 7.6: Illustration of findings in an image with their scores output by the detectors before (a), (c) and after (b), (d) inference by the CRF. The first row shows an example of an image that becomes more suspicious as a whole after reasoning about different findings, the second row shows an image that becomes less suspicious after inference.
Blue vertices represent soft tissue lesion candidates and red vertices microcalcification candidates. Solid vertices represent true positives.

comparison to local analysis is marginal, when evaluated at exam level. We suspect part of the reason for why the effect is small, is that in images where calcifications are present, these are often inside the patch that is used by the mass based CNN and the network already learns to associate calcifications with a higher risk. Even though the data is downscaled to 200 micron for the CNN pipeline and calcifications are typically difficult to see for humans at the resolution, the network may still find elements of these abnormalities. One of the parameters in the interaction potential learns several malignant findings close to each other are more likely to be normal. Even though we can see this effect in our validation and test set, the net effect it has on the exam based performance is minimal unfortunately. We suspect that when more training and test data becomes available, the interaction potential can have a more substantial effect.

In our model, we made use of three classes: normal, malignant mass and malignant calcification. However, more refined labels could be employed. For instance, as mentioned in the introduction, cysts are a common source of false positives and methods have been proposed to efficiently differentiate these from malignant masses [155]. By phrasing the mass classification as a three class problem, more detailed co-occurrence relations could be captured.

Geras et al. [97] also proposed a system that detects tumors in mammography that works on exams instead of regions of interest. To this end, they proposed a multi-stream CNN where all images are fed as a whole to the system. The system is not evaluated using biopsies as a reference and (F)ROC curves, but based on BIRADs categories and therefore results are difficult to compare with the method presented here and other mammography CAD papers. As we argued in the introduction, most of the task relevant information is in and around a patch centered at a potential tumor and therefore employing candidate detectors may be a more effective strategy.

An important design choice in a combination of systems is the point of fusion. For the classification of masses, we have made use of three unary CNN based potentials: (1) a CNN trained on a single (primary) patch (2) a CNN trained on both the primary and contra-lateral patch and (3) a CNN trained on the primary patch and a patch taken from the previous mammogram. Several other strategies could have been employed. As argued in Kooi et al. [151], all patches could be fed to a network simultaneously. This way, the model can potentially learn higher order interactions between all variables in the last fully connected layers. However, missing data can be an issue in this setting. In Kooi et al. [151] it was stipulated that for many temporal samples no prior is available. When no temporal samples are available, the potential can simply be set to $\psi(y_k) = (1, 1), \log[\psi(y_k)] = (0, 0)$, when employing the CRF described in this work. Similar principles apply to the addition of location information. In Kooi et al. [153], location features were added to the last fully connected layer and the system was retrained completely. In general, location information is available, but the training procedure becomes more complex again.

For the microcalcification pipeline, we have made use of a state-of-the art system, but this only takes into account the image that is being classified. Similar to the mass pipeline, more potential functions could be employed. Calcifications are often associated with ductal carcinoma in situ and can appear in a prior mammogram, but disappear in the current. By adding a temporal potential for the calcifications as well, this process could be modeled. Since the goal of this chapter is to present a fusion framework and no previous work has been published that investigates this, employing this is beyond the scope of this chapter.

When reading images, radiologists look at both the CC and MLO views for signs of cancer and need to reason about the image formation process and surrounding tissue, if something is only visible in one of the two views. Samulski et al. [227] presented a system that correlates candidate locations in the two standard views and uses several handcrafted features operating on the concatenation of the two regions. A similar approach to the symmetry and temporal networks used in this chapter and Kooi et. al [151] could be used to combine these two views, which could also be added as a potential function in the CRF model. Due to time constraints this was left to future work.

7.6 Conclusion

In this chapter, we employed a conditional random field (CRF) to integrate different sources of information from a mammographic exam and adopted the model to accommodate posteriors from different classifiers trained on different classes. We show that the model improves upon two state-of-the art mass and calcification detection systems that make use of only local information. We provided a method to determine exam based labels and evaluated two calibration methods to first scale the classifiers such that they represent accurate probabilities. The model provides a general framework

by formulating every source of information as a potential function, allowing for easy adjustment if new sources, such as additional patient covariates or other imaging modalities become available and for potentials to be trained discriminatively using popular models such as deep neural networks. Integration of information from multiple detections is a common problem in CAD and we believe the system not only has application in mammography but can also be used in other areas such as lung, retina and prostate CAD.

Chapter 8

Summary

Breast cancer is one of the most common types of cancer in the general population and most common cancer in women. In spite of advances in treatment it is still a leading cause of cancer death. Early detection has been shown to significantly improve chances of survival and therefore screening, where asymptomatic women in high risk subpopulations are invited for annual or biennial breast exams, is being performed in many countries. These programs generate an enormous amount of data which have to be read by trained experts, often a time consuming and error prone process. To mitigate this, computer aided detection and diagnosis (CAD) systems are being developed to aid and ultimately replace human readers of various medical images.

Until about 2010, most algorithms for image analysis tasks such as detection and segmentation were based on manually defined *features*: elements of the image engineers or medical professionals think are the most important. This changed with the introduction of *deep learning*: machine learning techniques that can learn which elements of a problem are important by providing a model with lots of raw data. These algorithms were initially applied to natural images and were received with skepticism in the medical image analysis community. The work in this thesis presents several new (deep) machine learning algorithms and frameworks for the detection of mass-like lesions in mammography and illustrates their power and potential for medical applications.

The first chapter provides an introduction to deep learning, background in training these models and their most important applications in medical image analysis and breast imaging. In chapter 2, a simple deep convolutional neural network (CNN) is compared to a state-of-the art CAD system and shown to outperform this system. The system relies on a candidate detector that pinpoints suspicious sites in an image, patches are extracted around these locations and presented to the CNN. An important part of the system is the data augmentation strategy we designed: patches are transformed to represent as many sources of variation that the system can encounter on test data, yet do not alter the class variable. This chapter illustrates the potential of deep neural networks in medical image analysis; with minimal engineering effort, state-of-the art performance can be obtained. The model is subsequently combined with several sets of traditional features, some of which convey information that is not fully captured by the CNN, such as the location or context of a potential malignant mass. The CNN is compared to a set of human readers and it was shown there is no significant difference in their performance.

Chapter 3 presents a feature based method to discriminate cysts from solid lesions in diagnostic mammography, that we developed before our work on deep neural networks. We analyzed sources of variation in the data, some of these are relevant for the classification problem and some only make learning more difficult. One of these nuisance factors is the amount of tissue surrounding a lesion. By developing two features that are invariant to this, the system is not presented with unnecessary information and the performance increases. In chapter 4, the same problem is tackled using a deep neural network. Rather than using features that are invariant, the dataset is transformed using data augmentation methods that simulate different amounts of occluding tissue. Next to each other, these chapters illustrate the shift in engineering effort that deep learning systems require: rather than spending time on developing features invariant to unnecessary information, the dataset is augmented with these sources of variation and the model is expected to learn an invariant representation from data.

Since medical images are typically too large to present to a network as a whole, models often work on patches, small

subregions of the image that readers of images or an automated system thinks are most suspicious. In the second chapter, we showed that a candidate detector can be used effectively to pinpoint these regions. Using such a system however, ignores any context information. Important sources of context in mammography are differences between left and right breast and temporal change: if the image under inspection has structures that are not in the contra lateral image or the prior, this can indicate the presence of a malignancy. Additionally, other markers for cancer such as calcifications are not explicitly modeled in the CNN. Chapters 5 and 6 presented methods to model this type of information.

In chapter 5, an algorithm for the detection of differences between left and right breast and temporal change is presented. Each patch used in the original CNN is mapped to a location in either the contra later or prior mammogram using a simple coordinate system of the breast. These two patches are subsequently fed to a multi stream CNN that learns if there are differences. We show that adding symmetry information can improve performance, but still find only marginal gains for temporal data and suspect better performance can be obtained when more data becomes available.

Chapter 6 presents a general framework to merge all the above mentioned methods into a system working on exams. The framework is based on a conditional random field, a statistical model that can be seen as an extension of a logistic regression to arbitrary sized inputs and outputs. The model learns co-occurrence relations between different mass findings and between mass and microcalcification findings. Additionally, several calibration methods to properly merge the systems are presented. We show that by combining symmetry, temporal, location and interaction information, a substantial improvement can be obtained upon the single patch system.

Chapter 9

General discussion

The 'godfather' of deep learning, Geoffrey Hinton, one of the most influential machine learning researchers of the past several decades, was recently asked what he thinks is the next big thing in deep learning. He replied:

> *Let me start by just saying a few things that seem obvious. I think if you work as a radiologist, you're like the coyote that's already over the edge of the cliff, but hasn't yet looked down so it doesn't realize there is no ground underneath him. People should stop training radiologists now, it is just completely obvious that within 5 years, deep learning is going to do better than radiologists, because it is going to be able to get a lot more experience. It may be 10 years, but we have plenty of radiologists already. I said this in a hospital and it didn't go down so well.*

Unfortunately, Hinton has not worked directly in radiology and seems to have only a vague idea of a radiologist's daily activities, though it does serve as a nice provocative statement and start of a discussion. For as far as I can see, it does contain elements of truth: many sub-tasks, in particular image reading, could be automated in the next (several) decade(s) and deep learning algorithms will likely play a key role. Other more high level tasks may follow in the decades after that and it is not unthinkable that somewhere in the future, large portions of a doctor's task may be replaced by a machine. Concern and excitement about artificial intelligence and potential replacement of (subtasks of) jobs in general is somewhat of a hot topic the past several years, especially since the introduction of efficient deep learning algorithms. At the moment there are roughly two streams:

1. People who point towards a major imminent shift in the job market [183], but are less worried about the general threat of AI. Frey and Osborne [87] provide a rigorous examination of future job prospects based on three computerization bottlenecks: *creativity*, *social intelligence* and *perception and manipulation* and conclude that nearly half of all jobs are at risk. Given recent advances in deep learning, people in the latter category may not be so safe anymore. In a recent survey, Grace and colleagues from Oxford [104] asked AI experts when and which jobs would be automated. 'AI researcher' turned out to be the last job AI researchers expected to be replaced. In these studies, algorithmic automation of white collar workers is often compared to the industrial revolution, where many blue collar workers were replaced by machines. The medical profession was still considered relatively safe.

2. People such as Nick Bostrom, whose ideas were popularized by Elon Musk and others who worry about the existence of our species [23] and safety of AI. To many people not directly involved in research on existential risk, this seems totally far fetched and like a scenario from a science fiction film, but serious institutes and researchers are spending time worrying about this. Oxford has a special department, called the Future of Humanity Institute [1] dedicated to existential risk problems and a similar institute backed by several MIT professors [2] was founded in Boston. The machine intelligence research institute [3] has a team of researchers dedicated to coming up with safe ways to harness AI. An open letter [4], signed by most prominent researchers in the field called for maximizing social benefit of AI and had an attached list of research priorities. The main argument researchers make about why we should worry now, is that humans often tend to make linear predictions about the future, but progress is exponential and it is therefore very difficult to estimate the impact AI will have in the -all but- near future.

[1] https://www.fhi.ox.ac.uk/
[2] https://futureoflife.org
[3] https://intelligence.org/about/
[4] https://futureoflife.org/ai-open-letter/

Rather than worrying, I think there is enough reason to be excited. When used safely, AI can really improve many people's lives. As computers take over subtasks, jobs will become more creative and working hours fewer, leaving more time to enjoy life. Doctors are generally intelligent and if automated systems were to replace parts of their activities, they could easily be retrained to spend a larger portion of their time on creative and research based work. For professions in general, more time may need to be spent on training and educating people to adjust as quickly as possible to new technology. During this project I experienced the rapid change in technology first hand. Although already widely applied in natural image analysis, deep learning was not accepted as a breakthrough technology in medical image analysis when I started this PhD and I consequently worked with a traditional system during the first year and a half. Subsequently, we had to make a switch and rapidly pick up new skills. Even during the development of DL algorithms, new and more efficient software and hardware started to emerge and code needed to be adapted many times over.

In my view, it is clear that deep learning technology brought about a paradigm shift in CAD research and transformed it from a 'very technical medical problem' to a 'machine learning problem' not requiring much medical knowledge. An illustrative example of this shift is the diabetic retinopathy challenge that was hosted by Kaggle [5]. Kaggle is an online platform where problems are posted in the form of challenges, by companies interested in having the problem solved. The winner is awarded a large sum of money and needs to make the approach and source code available. The diabetic retinopathy challenge revolved around the classification of stages of the disease from color fundus images. The winner, Ben Graham was a smart and well respected theoretical statistician, but had no previous exposure to the problem or medical imaging in general.

Notwithstanding limitations plain classification models have in comparison to human cognition and higher level tasks in radiology and medicine, I think no new technology needs to be developed to largely automate image reading and interpretation. Many problems revolve around image in-label out settings that can be solved by current technology: discriminative deep neural networks trained on large amounts of data. Although complaints about the limited amount of data in the medical domain are rife, for mammography this is definitely not the case. Some estimate roughly 100 million mammograms a year worldwide are recorded [141]. Actually automating image reading may be more of an organizational problem then a technical one: companies like Facebook and Google Deepmind have hundreds of the world top AI experts and software engineers in service but do not have access to the data yet. Hospitals or other institutes that do have large troves of data are unable to share it and do not have the budget or interest to hire AI experts like the big tech companies do. Though public datasets are slowly starting to emerge and excellent alternatives such as Kaggle challenges, the grand challenges hosted by DIAG, the DREAM challenges and Stanford's recently announced 'medical Imagenet' are appearing, few specialized AI groups are still working on medical problems.

Given their success in a wide range of AI problems, deep neural networks are recognized as the most powerful tool AI researchers currently have at their disposal. However, there remain a few shortcomings that limit their application and training efficiency.

9.1 Limitations of deep neural networks

Limitations can largely be divided into two categories: limitations in training data efficiency and the 'black box' nature of the models. Taking into account the current speed of development, some of the below mentioned limitations may already be solved by the time this thesis is printed.

9.1.1 One-shot and zero-shot learning

In terms of training efficiency, humans are still far superior to deep learning techniques. When humans see an object for the first time, we can generally recognize the object weeks if not years later under most lighting conditions and from most angles. Deep convolutional neural networks are currently trained by showing an image of an object and many of its different possible 2D representations that we expect to see during testing (data augmentation). On top of this, many different examples typically need to be provided before the network is able to recognize concepts. This is obviously not how humans do object recognition. Figure 9.1 shows an example of a chair that most people have never seen before but are likely to recognize as such, whereas deep neural networks will much harder time.

[5] www.kaggle.com

Figure 9.1: Example of a chair that most people have never seen before but are likely to recognize as such. Deep neural networks will have a much harder time (example taken from a talk by Max Welling).

Zero-shot learning[199] and one-shot [78] learning describe the ability of a machine learning system to learn to classify concepts that were omitted from the training set (zero-shot) and from a single example of a concept (one-shot). Several papers have also been published that derive deep neural network architectures with certain equivariant properties [61, 50] that we can use in combination with prior knowledge about the problem to reduce the amount of examples we need to present during training. Other more efficient training procedures that work with less data will likely be proposed. Generative models that make use of large amounts of unannotated data could play a big role in this.

9.1.2 Transfer learning and domain adaptation

Virtually all pilots are trained in flight simulators before flying a real plane and for good reason: this has been shown to improve their performance (and save expensive planes from crashing). In a similar spirit, people have been using networks trained on other data (the source data) and fine tuned for the task they are interested in solving (the target). For image analysis problems ImageNet is typically used to pre-train a model. The process of training a model to adapt to a novel problem is divided into two types [101]:

1. **Transfer learning:** in this setting the input data of the source and target task follow the same distribution, but the labels change. In a medical setting, this could entail the adaptation of a network trained to classify breast lesions into their respective BIRADS category from a network trained on normal vs. malignant labels.

2. **Domain adaptation:** in this setting the distribution of the source and target data change, but labels remain the same. This is also referred to as covariate shift adaptation [255] and can occur when, for instance, a network needs to be adapted to data from an institution with a different scanner where images have different properties.

Of course, both can occur simultaneously. When a model is transfered to a related task, the performance on the source task should improve as well. After all, if the flight simulator is designed properly, pilots will probably fly it better after flying in actual planes for some time. Currently, this is not the case. Machine learning algorithms adapt poorly to changing environments and once adapted, will likely perform the original task worse than after its original training. This phenomenon referred to as *catastrophic forgetting*. Though some recent work provided a solution for the forgetting of the first task [146], the performance on the first tasks does not increase after learning the second problem.

9.1.3 Competence without comprehension

Evolutionary computing, neural networks and other nature inspired algorithms are beautiful because they can be used to create machines that solve problems we do not know how to solve ourselves. However, this 'competence without comprehension' [57] comes at a price: explaining to the end-user how a certain decision came about is a big problem and will limit the application of current deep neural network architectures [66, 168]. Starting from 2018, the European union (EU) will put laws in place that

> [...] will restrict automated individual decision-making (that is, algorithms that make decisions based on user level predictors) which "significantly affect" users. The law will also effectively create a "right to explanation," whereby a user can ask for an explanation of an algorithmic decision that was made about them. [103]

These laws may have serious implications for commercial implementations of deep neural nets in health care, because explaining why a neural net identifies a region in an image as cancerous, for instance, is not easy. Similarly, fintech companies working on applications such as fraud detection are currently struggling to sell deep learning based systems, due to laws that require companies to explain why a person or transactions was identified as fraudulent, that are already in place in the US.

DARPA recently announced the explainable AI program (XAI) [6] dedicated to making models interpretable and other groups are already experimenting with potential solutions. One such attempt is a model that, apart from a distribution over possible labels, outputs a natural sentence that explains why the model derived the output [201]. Other interesting ideas revolve around training a simpler, more explicable model with the output of the deep model. It turns out there is so much information in these labels, that you can approximate a convolutional neural network quite accurately with a simple model such as a logistic regression, that has more interpretable weights [117]. For models using raw signals as input this will not solve much unfortunately. Proving that the natural sentence or human interpreting the weights in the logistic regression is actually what the decision is based on is, of course, difficult. But then again, it may be just as difficult to prove the equivalence of the factors we think we base our decisions on and those actually involved in the decision making process. There is substantial evidence much of our decision making is subconscious. For instance, by showing a person the word 'apple' several times before asking them if they want an apple or pear, they are more likely to choose the first, though most people will make up a story about why they choose it.

9.1.4 Adversarial images

Somewhat related to the inexplicable 'black box' nature of deep neural networks as outlined above is the case of *adversarial examples*. It turns out that deep neural networks are extremely brittle when it comes to signals that are generated specifically to fool the model. By evaluating the output, the input can be changed in such a way that the network optimally makes the 'wrong' decision, akin to how the parameters in the model are learned. This is particularly troublesome in security applications. Identity theft can take place by generating fingerprints, irises or faces that resemble the target person very accurately, according to the system in place. Figure 9.2 provides several examples of objects that were made to look like an ostrich (to the network) by adding noise that is (nearly) imperceptible to humans.
Though security is less of a problem in medical image analysis, adversarial-like examples are present and can seriously disturb a system. During the development of the methods described in this thesis, I have encountered numerous examples of abnormal, but benign tissue that would never have been identified as malignant by a radiologist. These are, however, not specifically designed to fool the network but just deviate so much from examples in the training data that they land on the wrong side of the discriminant. For research projects this is not a major issue, but before clinical application as an independent system can take place, these issues need to be resolved.

9.2 Limitations of the current system

All algorithms in this thesis were designed with the goal to ultimately automate the interpretation of screening mammograms. Although results are promising and certain subproblems are more or less solved, there is still work that needs to be done and several shortcomings of the current system can be identified. If a similar approach is followed, where regions of interest are classified instead of whole images or cases, the following elements could be improved:

[6] https://www.darpa.mil/program/explainable-artificial-intelligence

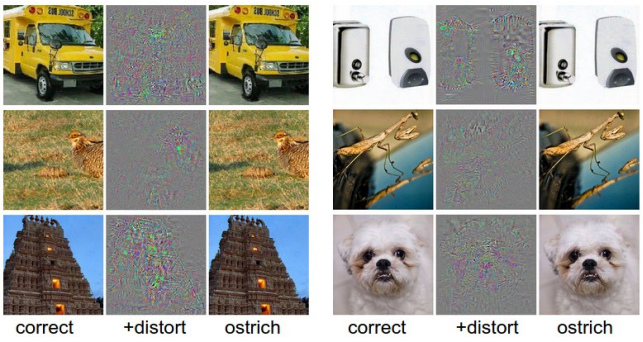


Figure 9.2: Example of adversarial images specifically designed to fool a neural network. Every left image is classified correctly, the center image is the noise added to the image designed to make the network think the image is an ostrich. (Source: Intriguing properties of neural networks - Szegedy et al. 2014 [261]).

- **Temporal analysis:** We know from reader studies that comparing current and prior exam results in a significant increase in specificity of readers [267, 32, 280, 216, 291]. In chapters 5 and 6, we have added temporal information, which gave minor improvements in performance, but were marginal in comparison to what one would expect from these studies. Similarly, although an increase in performance was observed in earlier work using traditional features [268, 269], this increase was not as great as one would expect given the results from reader studies.

- **Calcification detection:** The detection of calcifications remains a hard problem and the system used in this paper does not operate at a human level yet. The problem is, in my view, substantially more difficult than classifying soft tissue lesions because individual calcifications are small, whereas large windows are required for classifying groups or clusters. Since the calcifications are small, (naively) downscaling the image is not an option.

- **CC/MLO correlation:** When reading images, radiologists look at both the CC and MLO views for signs of cancer and need to reason if something is only visible in one of the two views. Samulski et al. [227] presented a system that maps locations in the two views and uses several handcrafted features operating on the concatenation of the two regions. A similar approach to the symmetry and temporal networks presented in chapter 5 of this thesis could be used to combine these two views and could also be added as a potential function in the CRF model presented in chapter 6.

- **Context of the entire mammogram**: Hupse et al. [125] defined several handcrafted features summarizing the context in several regions of the breast. These features were added to the last fully connected layer of a simple CNN and shown to improve performance. As discussed in 2.1.4, multistream network architectures are often used to perform multi-scale image analysis. A similar approach that takes all the context of the mammogram into account could be used in the current system. Elements such as breast density are known to be risk factors and could be captured like this.

Estimating the potential contribution of each of these elements to improving the system is a complicated task though. In the study presented in chapter 2, we compared the CNN to a set of expert readers that were asked to classify regions of interest and showed the performance of our system is comparable to these readers on a patch level. We also showed that when combining the scores of these readers, the performance significantly improves. Similar results were obtained by Karssemeijer et al. [139] who combined the scores of twelve radiologist for whole cases and showed the performance keeps increasing, but unfortunately no studies exist for all these individual factors.

Not all tasks can be learned from any type of data. For instance, we can not learn to discriminate red cars from blue cars by only looking at gray scale images (if the dataset is unbiased). These types of reader studies are interesting because they provide a lower bound on the maximum amount of information in the test set: even if we reach human performance, an algorithm could still perform better. However, this does not mean this performance can be obtained with any model and any training data. As I also argued in the introduction, there is enough reason to believe that the goal of greatly outperforming human readers on most of these tasks is realistic and I think this is possible in the next 5 years. Other, more elegant and simple models that work directly on cases (collection of images) and bypass the whole stage-wise approach presented in this thesis could be developed for this.

9.3 Future directions

9.3.1 Reducing data scarcity

Medical data is far from scarce. However, well curated, annotated public data still are, unfortunately. The machine learning community has been working on ways to alleviate the need for large datasets of finely annotated data for decades and some methods have recently found their way to medical image analysis. Two main trends can be distinguished: (1) methods that work with coarser annotations, thereby alleviating the burden to outline every finding or annotate every sample in a dataset and (2) methods that stimulate sharing of medical data or facilitate training of models without explicit sharing.

- **Coarse annotations**
  - *Semi-supervised learning*
    In the introduction, I described the commonly used distinction between supervised and unsupervised learning. In supervised learning, the parameters of a model are fit to a dataset $\mathcal{D} = \{\mathbf{x}_n, y_n\}_{n=1}^N$ of input $\mathbf{x}$ and output y examples. In the past several decades, many variations on these concepts have been explored. Semi-supervised learning is one of these and refers to the concept where the parameters of the model are fit to two datasets, labeled data $\mathcal{D}_L = \{\mathbf{x}_n, y_n\}_{n=1}^N$ and unlabeled data $\mathcal{D}_U = \{\mathbf{x}_n\}_{n=1}^N$, where $\mathcal{D}_L$ and $\mathcal{D}_U$ are generally assumed to be sampled from the same data-generating function. A simple example of a semi-supervised learning algorithm, is to use a clustering method to group data based on similar characteristics and assign each group the most occurring label of the annotated samples in the particular group.

    Using this concept, massive amounts of unannotated data can be used to improve the performance of learning algorithms. Additionally, the concept could be used in an 'online' setting. Imagine a product running some classifier on incoming, unannotated data. Each scan from every customer the system classifies is unlabeled, but can be added to the training data to improve the performance of the system. Related to this concept is *active learning*: a paradigm where algorithms query annotators for labels and try to get the optimal amount of information out of the novel labels, thereby minimizing annotation effort.
  - *Weakly labeled data*
    In this thesis, I have worked with data that contained annotations in the form of contours around lesions or microcalcifications in images. We can see this, somewhat ad-hoc, as an {image + contour-label}-in $\rightarrow$ {image-label}-out setting, where the input is the training phase and output the prediction phase. Annotating contours is time consuming and in our case unnecessary, since CNNs simply work with patches. Rough centers of lesions would, therefore, be sufficient input.

    An even faster way to annotate data, would be to provide labels on an image level rather than for every finding in the scan. This effectively removes information and is especially complicated when multiple diseases are present in an image, but can reduce annotation time of large datasets. Working with annotations on an image, rather than bounding box or contour level is referred to as *weakly labeled* learning. For clarity, this is different, though closely related to the slightly more complicated task of *multiple instance learning* (MIL). Working with weakly labeled data can be seen as an {image, image-label}-in $\rightarrow$ {image-label}-out system and MIL as an {image, image-label}-in $\rightarrow$ {contour-label}-out system.
  - *Unstructured labels*
    The idea of working with weakly labeled images can be taken one step further by simply not using labels at all anymore, but instead work with unstructured natural language labels provided by doctors. Whether this removes or adds information is difficult to say: contours provide exact delineations and possibly the joint knowledge of a radiology report and an experienced annotator. A full report may contain some additional information not conveyed by the contour and also carry the sentiment of the radiologist. In any case, working with this effectively renders the annotator unnecessary and systems learning from natural sentences could save time and money.

  Interesting studies would investigate the amount of data that is needed to obtain a certain performance with all these types of annotations. This way, the benefits and cost of annotation can be carefully considered before starting a new project.

- **Data sharing**

– *Differential privacy and privacy preserving deep learning* As mentioned above, a potential downside of deep learning methods is that large troves of annotated training data are typically required to make the models work properly. Institutions are reluctant or unable to share their data, some to maintain a competitive edge over other institutes, but some also purely because of privacy laws. Even if data is shared and properly anonymized, it could still reveal sensitive information. Although deep neural networks are difficult to interpret, there is evidence that at least part of the training data can be traced back by something called a 'model-inversion' attack [83], even if the attacker only has access to the model's in- and output. This information can, for instance, stem from overfitting, where the model memorizes part of the training data. This has been shown to expose a patient's identity in a genomics project [84].

  Differential privacy [68, 133, 2] algorithms constitute methods that aggregate statistics about datasets, revealing as little information as possible about the identity of individual samples. Building on these concepts, methods have been proposed [237] that distribute the training of deep neural networks and ensure privacy is preserved. This way, deep neural networks can be trained on small datasets from individual institutions and aggregated in a center controlling the training process, without actually sharing any data explicitly. This concept is referred to as *private-multiparty machine learning* and had a dedicated workshop at a major machine learning conference in 2016 [7].

– *Block chain technology* Block chain technology, a revolutionary new technology blazoned as the internet of the 21st century, lies at the foundation of the recently emerging cryptocurrencies such as Bitcoin and Ethereum that are taking the world by storm, at the time of writing. It relies on an immutable and distributed ledger, an electronic record of transactions, where new transactions are appended and shared with all copies of the ledger that is owned by everyone and no-one. Hacking of medical data is rife and medical records sell for sometimes orders of magnitude more than credit card details. Additionally, patients have little control over their data.

  Using block chain technology, electronic health records (EHR) go into a ledger and are safely transmitted to other institutions [8]. This way, patients will have more control over their data with improved security and ease of access for research institutes. The MedRec system [7] is a prototype developed by the MIT media lab relying on the Ethereum [33] block chain. The miners of the system are medical researches, who are rewarded with data instead of digital currency. This technology is already implemented in the Bowhead system [9], which analyzes blood and saliva samples that patients can sell anonymously and transmit through a blockchain to research institutes.

9.3.2 Different types of CAD

- **Multi-class CAD** The work in this thesis made use of binary labels only: tumor or no tumor, but more can be done. It is very straightforward to extend CNNs to perform multi-class classification and work with more detailed labels such as malignant mass, asymmetry, architectural distortion, cyst, other normal tissue, etc. By providing the exact label and not simply normal/malignant, the entropy decreases and this should lead to better performance when trained properly.

- **Multi-label CAD** Multiple diseases can occur simultaneously in some images, but most CAD systems are currently designed to look for a single specific type of abnormality or one specific type of disease. As stipulated in the introduction, mammography is almost exclusively used for the detection of breast cancer, but breast arterial calcifications (BACs) can also be seen on mammograms and have been associated with heart disease. Hence, CAD could be extended to analyze multiple types of abnormalities.

- **Multi-modal CAD** Though less common in screening, in a diagnostic setting different types of images of a patient are typically recorded and it is likely that these images will be complementary to some extent. For instance, as mentioned in the introduction, women in high risk populations for breast cancer will receive an additional MRI. Since microcalcifications are not visible on MRI, but can be seen on a mammogram it makes sense to work towards

[7] https://pmpml.github.io/PMPML16/
[8] https://www.youtube.com/watch?v=GO9Q7i-IcA8
[9] https://bowheadhealth.com/

a system that takes both into account and generates a single score for the entire structure. Combining only the outputs of different CNNs for the system will ignore any correlation between image values of the different modalities.

These general systems can go beyond images alone. In many settings, far more information about a patient is available and all these covariates can be taken into account. AI methods for electronic health records are emerging [228, 202] and are very exciting. These could be combined with information in images. Genomic test such as the Oncotype DX and Mammaprint analyze a subset of genes associated with breast cancer and could be used in conjunction with imaging data to make more accurate diagnosis. Blood tests have also been developed to diagnose cancer. For instance, Thomas Wurdinger's group developed a method that could diagnose the presence of localized and metastasized tumors with 96% accuracy and the location of six different tumor types with 71% accuracy [19].

In the future, this can go much further and Internet of things (IoT)-like setting for health are likely to emerge. Chips in the toilet can analyze stool and urine samples, chips under your skin analyze blood values 24/7, chips in the sink or a tooth brush analyze saliva, other chips analyze sweat and breath. This data can be combined with data from fitness trackers, smart phones and genetic information into a single system that provides food/exercise/sleep recommendations at will and rings an alarm when anomalies are detected. If all this data is safely accumulated and stored in a world wide health database, data science can make massive advancements. The ShenZhen based startup ICarbonX [10] has the ambitious goal to merge all this information in a single system and similar efforts being made at IBM Watson's health department.

- **Computer aided treatment planning** Most machine learning done in radiology, opthalmology and pathology so far revolves around assigning a label to an image or a set of pixels in the image. However, doctors are subjected to more complex problems for which machine learning can also be used. Treatment planning is one of those. One way to phrase this is in the form of a sequential optimization problem. A system receives observations of a patient, which correspond to their internal states and actions the expert has to choose from, corresponding to admissible treatment strategies. There can be uncertainty in the states of the patients, i.e., there is a stochastic process governing the observations given the states and there can also be uncertainty in the outcome of every administered treatment. An algorithm can be used to provide the optimal action based on an observation and transition model. These types of decision problems are know as Markov decision processes (MDP) when there is no uncertainty in observations and partially observable Markov decision process (POMDP) when the underlying state is unobserved.

  The parameters of the observation and transition model can also be learned from data. This type of learning is known as *reinforcement learning*, which is another recently reinvigorated type of machine learning and is the key ingredient to Deep Mind's Atari game playing software [185] and AlphaGo [243], the system that recently beat world champion Lee Se-Dol. No medical applications exist yet, but prototypes have been proposed based on simulated data [73].

[10] https://www.icarbonx.com/

Bibliography

[1] Martın Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv:1603.04467*, 2016.

[2] Martín Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318. ACM, 2016.

[3] Ayelet Akselrod-Ballin, Leonid Karlinsky, Sharon Alpert, Sharbell Hasoul, Rami Ben-Ari, and Ella Barkan. A region based convolutional network for tumor detection and classification in breast mammography. In *DLMIA*, volume 10008 of *Lect Notes Comput Sci*, pages 197–205, 2016.

[4] Paul D Allison. *MIssing Data*. Sage Publications, 2001.

[5] John Arevalo, Fabio A González, Raul Ramos-Pollán, Jose L Oliveira, and Miguel Angel Guevara Lopez. Representation learning for mammography mass lesion classification with convolutional neural networks. *Comput Methods Programs Biomed*, 127:248–257, 2016.

[6] S. M. Astley and F. J. Gilbert. Computer-aided detection in mammography. *Clin Radiol*, 59:390–399, 2004.

[7] Asaph Azaria, Ariel Ekblaw, Thiago Vieira, and Andrew Lippman. Medrec: Using blockchain for medical data access and permission management. In *Open and Big Data (OBD), International Conference on*, pages 25–30. IEEE, 2016.

[8] Yaniv Bar, Idit Diamant, Lior Wolf, and Hayit Greenspan. Deep learning with non-medical training used for chest pathology identification. In *Medical Imaging*, volume 9414 of *Proceedings of the SPIE*, page 94140V, 2015.

[9] Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian Goodfellow, Arnaud Bergeron, Nicolas Bouchard, David Warde-Farley, and Yoshua Bengio. Theano: new features and speed improvements. In *Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop*, 2012.

[10] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Trans Pattern Anal Mach Intell*, 35(8):1798–1828, 2013.

[11] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. In *Advances in Neural Information Processing Systems*, pages 153–160, 2007.

[12] Yoshua Bengio. Learning deep architectures for ai. *Foundations and trends in Machine Learning*, 2:1–127, 2009.

[13] Yoshua Bengio. Practical recommendations for gradient-based training of deep architectures. In *Neural Networks: Tricks of the Trade*, pages 437–478. Springer Berlin Heidelberg, 2012.

[14] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans Neural Netw*, 5:157–166, 1994.

[15] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *J Mach Learn Res*, 13(1):281–305, 2012.

[16] James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*, volume 4, page 3, 2010.

[17] Eta S Berner and Mark L Graber. Overconfidence as a cause of diagnostic error in medicine. *The American journal of medicine*, 121(5):S2–S23, 2008.

[18] Julian Besag. Efficiency of pseudolikelihood estimation for simple gaussian fields. *Biometrika*, 1977.

[19] Myron G Best, Nik Sol, Irsan Kooi, Jihane Tannous, Bart A Westerman, François Rustenburg, Pepijn Schellen, Heleen Verschueren, Edward Post, Jan Koster, et al. Rna-seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics. *Cancer cell*, 28(5):666–676, 2015.

[20] Archie Bleyer and H Gilbert Welch. Effect of three decades of screening mammography on breast-cancer incidence. *New England Journal of Medicine*, 367(21):1998–2005, 2012.

[21] Hans Bornefalk and Anna Bornefalk Hermansson. On the comparison of froc curves in mammography CAD systems. *Med Phys*, 32:412–417, 2005.

[22] Brian H Bornstein and A Christine Emler. Rationality in medical decision making: a review of the literature on doctors' decision-making biases. *Journal of evaluation in clinical practice*, 7(2):97–107, 2001.

[23] Nick Bostrom. *Superintelligence: Paths, dangers, strategies*. OUP Oxford, 2014.

[24] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[25] A. Bria, N. Karssemeijer, and F. Tortorella. Learning from unbalanced data: A cascade-based approach for detecting clustered microcalcifications. *Med Image Anal*, 18:241–252, 2013.

[26] John Brodersen and Volkert Dirk Siersma. Long-term psychosocial consequences of false-positive screening mammography. *The Annals of Family Medicine*, 112(2):106–115, 2013.

[27] Mireille Broeders, Sue Moss, Lennarth Nyström, Sisse Njor, Hkan Jonsson, Ellen Paap, Nathalie Massat, Stephen Duffy, Elsebeth Lynge, and Eugenio Paci. The impact of mammographic screening on breast cancer mortality in europe: a review of observational studies. *J Med Screen*, 19:1425, 2012.

[28] Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. 35(8):1872–1886, 2013.

[29] Bruce G Buchanan, Edward Hance Shortliffe, et al. *Rule-based expert systems*, volume 3. Addison-wesley Reading, MA, 1984.

[30] P Burlina, D E Freund, N Joshi, Y Wolfson, and N M Bressler. Detection of age-related macular degeneration via deep learning. In *IEEE Int Symp Biomedical Imaging*, pages 184–188, 2016.

[31] Elizabeth S Burnside, Jesse Davis, Jagpreet Chhatwal, Oguzhan Alagoz, Mary J Lindstrom, Berta M Geller, Benjamin Littenberg, Katherine A Shaffer, Charles E Kahn Jr, and C David Page. Probabilistic computer model developed from clinical data in national mammography database format to classify mammographic findings. *Radiology*, 251(3):663–672, 2009.

[32] Elizabeth S Burnside, Edward A Sickles, Rita E Sohlich, and Katherine E Dee. Differential value of comparison with previous examinations in diagnostic versus screening mammography. *American Journal of Roentgenology*, 179(5):1173–1177, 2002.

[33] Vitalik Buterin et al. A next-generation smart contract and decentralized application platform. *white paper*, 2014.

[34] Miguel A Carreira-Perpinan and Geoffrey Hinton. On contrastive divergence learning. In *AISTATS*, volume 10, pages 33–40, 2005.

[35] Paola Casti, Arianna Mencattini, Marcello Salmeri, and Rangaraj M Rangayyan. Analysis of structural similarity in mammograms for detection of bilateral asymmetry. 34(2), 2015.

[36] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. In *arXiv:1606.01865*, 2016.

[37] Cristina M. Checka, Jennifer E. Chun, Freya R. Schnabel, Jiyon Lee, and Hildegard Toth. The relationship of mammographic density and age: implications for breast cancer screening. *AJR Am J Roentgenol*, 198:W292–W295, 2012.

[38] Hao Chen, Dong Ni, Jing Qin, Shengli Li, Xin Yang, Tianfu Wang, and Pheng Ann Heng. Standard plane localization in fetal ultrasound via domain transferred deep neural networks. *IEEE J Biomed Health Inform*, 19(5):1627–1636, 2015.

[39] Hao Chen, Xiaojuan Qi, Lequan Yu, and Pheng-Ann Heng. DCAN: Deep contour-aware networks for accurate gland segmentation. *Med Image Anal*, 36:135–146, 2017.

[40] Liang-Chieh Chen, Alexander G Schwing, Alan L Yuille, and Raquel Urtasun. Learning deep structured models. *arXiv:1407.2538*, 2014.

[41] Sihong Chen, Jing Qin, Xing Ji, Baiying Lei, Tianfu Wang, Dong Ni, and Jie-Zhi Cheng. Automatic scoring of multiple semantic attributes with multi-task feature leverage: A study on pulmonary nodules in CT images. *IEEE Trans Med Imaging*, 2016. , in press.

[42] S. C. Cheng and Y. M. Huang. A novel approach to diagnose diabetes based on the fractal characteristics of retinal images. *IEEE Trans Inf Technol Biomed*, 7:163–170, 2003.

[43] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv:1406.1078*, 2014.

[44] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Med Image Comput Comput Assist Interv*, volume 9901 of *Lect Notes Comput Sci*, pages 424–432. Springer, 2016.

[45] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, and Andrea Vedaldi. Deep filter banks for texture recognition, description, and segmentation. *Int J Comput Vis*, 118:65–94, 2016.

[46] Francesco Ciompi, Bartjan de Hoop, Sarah J. van Riel, Kaman Chung, Ernst Th. Scholten, Matthijs Oudkerk, Pim A de Jong, Mathias Prokop, and Bram van Ginneken. Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box. *Med Image Anal*, 26:195–202, 2015.

[47] Dan C. Cireşan, Alessandro Giusti, Luca M. Gambardella, and Jürgen Schmidhuber. Mitosis detection in breast cancer histology images with deep neural networks. In *Med Image Comput Comput Assist Interv*, volume 8150 of *Lect Notes Comput Sci*, pages 411–418, 2013.

[48] Dan C. Cireşan, Ueli Meier, Jonathan Masci, and Jürgen Schmidhuber. Multi-column deep neural network for traffic sign classification. *Neural Netw*, 32:333–338, 2012.

[49] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). 2015.

[50] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning*, pages 2990–2999, 2016.

[51] Taco S Cohen and Max Welling. Transformation properties of learned visual representations. *arXiv:1412.7659*, 2014.

[52] Ronan Collobert, Koray Kavukcuoglu, and Clément Farabet. Torch7: A matlab-like environment for machine learning. In *Advances in Neural Information Processing Systems*, 2011.

[53] Pat Croskerry. The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic medicine*, 78(8):775–780, 2003.

[54] M.U. Dalmis, G. Litjens, K. Holland, A. Setio, R. Mann, N. Karssemeijer, and A. Gubern-Mérida. Using deep learning to segment breast and fibroglandular tissue in mri volumes. *Medical physics*, 44:533–546, February 2017.

[55] Yann N Dauphin, Harm de Vries, Junyoung Chung, and Yoshua Bengio. Rmsprop and equilibrated adaptive learning rates for non-convex optimization. *arXiv:150204390*, 2015.

[56] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[57] Daniel C. Dennett. *From Bacteria to Bach and Back: The Evolution of Minds*. W. W. Norton, 2017.

[58] Carol E DeSantis, Chun Chieh Lin, Angela B Mariotto, Rebecca L Siegel, Kevin D Stein, Joan L Kramer, Rick Alteri, Anthony S Robbins, and Ahmedin Jemal. Cancer treatment and survivorship statistics, 2014. *CA Cancer J Clin*, (4):252–271, 2014.

[59] Neeraj Dhungel, Gustavo Carneiro, and Andrew P Bradley. The automated learning of deep features for breast mass classification from mammograms. In *Med Image Comput Comput Assist Interv*, volume 9901 of *Lect Notes Comput Sci*, pages 106–114. Springer, 2016.

[60] Sander Dieleman, Jeffrey De Fauw, and Koray Kavukcuoglu. Exploiting cyclic symmetry in convolutional neural networks. *arXiv:1602.02660*, 2016.

[61] Sander Dieleman, Kyle W Willett, and Joni Dambre. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly notices of the royal astronomical society*, 450(2):1441–1459, 2015.

[62] John Michael Dixon, C McDonald, Robert Elton, and William Miller. Risk of breast cancer in women with palpable breast cysts: a prospective study. *Lancet*, 353(9166):1742–1745, 1999.

[63] K. Doi. Current status and future potential of computer-aided diagnosis in medical imaging. *Br J Radiol*, 78 Spec No 1:S3–S19, 2005.

[64] Kunio Doi. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Comput Med Imaging Graph*, 31:198–211, 2007. PMID: 17349778.

[65] Pedro Domingos. *The Master Algorithm*. Basic Books, 2016.

[66] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv:1702.08608*, 2017.

[67] Anastasia Dubrovina, Pavel Kisilev, Boris Ginsburg, Sharbell Hashoul, and Ron Kimmel. Computational mammography using deep neural networks. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, pages 1–5, 2016.

[68] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[69] Bradley Efron. Bootstrap methods: Another look at the jackknife. *Ann Stat*, 7(1):1–26, 1979.

[70] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*, volume 57. CRC press, 1994.

[71] M. Elter and A. Horsch. Cadx of mammographic masses and clustered microcalcifications: a review. *Med Phys*, 36:2052–2068, 2009.

[72] Klaus Erhard, Fleur Kilburn-Toppin, Paula Willsher, Elin Moa, Erik Fredenberg, Nataly Wieberneit, Thomas Buelow, and Matthew G Wallis. Characterization of cystic lesions by spectral mammography: Results of a clinical pilot study. *Investigative Radiology*, 51(5):340–347, 2016.

[73] Damien Ernst, Guy-Bart Stan, Jorge Goncalves, and Louis Wehenkel. Clinical data based optimal sti strategies for hiv: a reinforcement learning approach. In *Decision and Control, 2006 45th IEEE Conference on*, pages 667–672. IEEE, 2006.

[74] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118, 2017.

[75] Karla K. Evans, Robyn L. Birdwell, and Jeremy M. Wolfe. If you don't find it often, you often don't find it: why some cancers are missed in breast cancer screening. *PLoS One*, 8:e64366, 2013.

[76] Karla K Evans, Robyn L Birdwell, and Jeremy M Wolfe. If you don't find it often, you often don't find it: why some cancers are missed in breast cancer screening. *PLoS One*, 8(5):e64366, 2013.

[77] Clément Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *IEEE Trans Pattern Anal Mach Intell*, 35(8):1915–1929, 2013.

[78] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.

[79] Joshua J Fenton, Linn Abraham, Stephen H Taplin, Berta M Geller, Patricia A Carney, Carl D'Orsi, Joann G Elmore, and William E Barlow. Effectiveness of computer-aided detection in community mammography practice. *J Natl Cancer Inst*, 103:1152–1161, 2011.

[80] R. J. Ferrari, R. M. Rangayyan, J. E. Desautels, and A. F. Frre. Analysis of asymmetry in mammograms via directional filtering with gabor wavelets. *IEEE Trans Med Imaging*, 20:953–964, 2001.

[81] P Fonseca, J. Mendoza, J. Wainer, J. Ferrer, J. Pinto, and B. Guerrero, J.and Castaneda. Automatic breast density classification using a convolutional neural network architecture search procedure. In *Medical Imaging*, volume 9413 of *Proceedings of the SPIE*, page 941428, 2015.

[82] Sergei V. Fotin, Yin Yin, Hrishikesh Haldankar, Jeffrey W. Hoffmeister, and Senthil Periaswamy. Detection of soft tissue densities from digital breast tomosynthesis: comparison of conventional and deep learning approaches. In *Medical Imaging*, volume 9785 of *Proceedings of the SPIE*, page 97850X, 2016.

[83] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333. ACM, 2015.

[84] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *Proceedings of the 23rd USENIX Security Symposium.*, 2014.

[85] T. W. Freer and M. J. Ulissey. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. *Radiology*, 220:781–786, 2001.

[86] Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer, 1995.

[87] Carl Benedikt Frey and Michael A Osborne. The future of employment: how susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, 114:254–280, 2017.

[88] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.

[89] Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. *Machine learning*, 29:131–163, 1997.

[90] K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol Cybern*, 36(4):193–202, 1980.

[91] Brian Fulkerson, Andrea Vedaldi, and Stefano Soatto. Class segmentation and object localization with superpixel neighborhoods. In *Int Conf Comp Vis*, 2009.

[92] Carolina Galleguillos, Brian McFee, Serge Belongie, and Gert Lanckriet. Multi-class object localization by combining local contextual interactions. In *Comput Vis Pattern Recognit*, 2010.

[93] Mingchen Gao, Ziyue Xu, Le Lu, I Nogues, R Summers, and D Mollura. Segmentation label propagation using deep convolutional neural networks and dense conditional random field. In *IEEE Int Symp Biomedical Imaging*, pages 1265–1268, 2016.

[94] Y Gao, Mohammad Ali Maraci, and J Alison Noble. Describing ultrasound video content using deep convolutional neural networks. In *IEEE Int Symp Biomedical Imaging*, pages 787–790, 2016.

[95] Shantanu Gaur, Vandana Dialani, Priscilla J Slanetz, and Ronald L Eisenberg. Architectural distortion of the breast. *American Journal of Roentgenology*, 201(5):W662–W670, 2013.

[96] Robert Gens and Pedro M Domingos. Deep symmetry networks. In *Advances in Neural Information Processing Systems*, pages 2537–2545, 2014.

[97] Krzysztof J Geras, Stacey Wolfson, S Kim, Linda Moy, and Kyunghyun Cho. High-resolution breast cancer screening with multi-view deep convolutional neural networks. *arXiv:1703.07047*, 2017.

[98] J-M Geusebroek, Rein van den Boomgaard, Arnold W. M. Smeulders, and Hugo Geerts. Color invariance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(12):1338–1350, 2001.

[99] M. L. Giger, N. Karssemeijer, and S. G. Armato. Computer-aided diagnosis in medical imaging. *IEEE Trans Med Imaging*, 20:1205–1208, 2001.

[100] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jagannath Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Comput Vis Pattern Recognit*, pages 580–587. IEEE, 2014.

[101] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT Press, 2016.

[102] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *arXiv:1406.2661*, 2014.

[103] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a" right to explanation". *arXiv:1606.08813*, 2016.

[104] Katja Grace, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. When will ai exceed human performance? evidence from ai experts. *arXiv:1705.08807*, 2017.

[105] Alex Graves. *Supervised Sequence Labelling with Recurrent Neural Networks*. PhD thesis, 2012.

[106] H. Greenspan, R. M. Summers, and B. van Ginneken. Deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Trans Med Imaging*, 35(5):1153–1159, 2016.

[107] Klaus Greff, Rupesh Kumar Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. *arXiv:1503.04069*, 2015.

[108] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C Nelson, Jessica L Mega, and Dale R Webster. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316:2402–2410, December 2016.

[109] L. Hadjiiski, B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. Gurcan. Analysis of temporal changes of mammographic features: computer-aided classification of malignant and benign breast masses. *Med Phys*, 28:2309–2317, 2001.

[110] Robert M. Haralick, K. Shanmugam, and Its'Hak Dinstein. Textural features for image classification. *IEEE Trans Syst Man Cybern*, 3:610–621, 1973.

[111] John Haugeland. *Artificial intelligence: the very idea*. The MIT Press, Cambridge, Mass, 1985.

[112] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv:1512.03385*, 2015.

[113] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Comput Vis Pattern Recognit*, pages 1026–1034, 2015.

[114] D. Heckerman and J. S. Breese. Causal independence for probability assessment and inference using Bayesian networks. *IEEE Trans Syst Man Cybern*, 26:826–831, 1996.

[115] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313:504–507, 2006.

[116] Geoffrey Hinton. A practical guide to training restricted Boltzmann machines. *Momentum*, 9(1):926, 2010.

[117] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv:1503.02531*, 2015.

[118] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

[119] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Comput*, 18:1527–1554, 2006.

[120] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[121] Johannes Hofmanninger and Georg Langs. Mapping visual features to semantic profiles for retrieval in medical imaging. In *Comput Vis Pattern Recognit*, pages 457–465, 2015.

[122] Jan Hosang, Rodrigo Benenson, Piotr Dollár, and Bernt Schiele. What makes for effective detection proposals? *arXiv:150205082*, 2015.

[123] Ming-Kuei Hu. Visual pattern recognition by moment invariants. *Information Theory, IRE Transactions on*, 8(2):179–187, 1962.

[124] R. Hupse and N. Karssemeijer. Use of normal tissue context in computer-aided detection of masses in mammograms. *IEEE Trans Med Imaging*, 28:2033–2041, 2009.

[125] Rianne Hupse and Nico Karssemeijer. The use of contextual information for computer aided detection of masses in mammograms. In *Medical Imaging*, volume 7260 of *Proceedings of the SPIE*, page 72600Q, 2009.

[126] Rianne Hupse, Maurice Samulski, Marc Lobbes, Ard den Heeten, Mechli W. Imhof-Tas, David Beijerinck, Ruud Pijnappel, Carla Boetes, and Nico Karssemeijer. Standalone computer-aided detection compared to radiologists' performance for the detection of mammographic masses. *Eur Radiol*, 23:93–100, 2013.

[127] Rianne Hupse, Maurice Samulski, Marc B. Lobbes, Ritse M. Mann, Roel Mus, Gerard J. den Heeten, David Beijerinck, Ruud M. Pijnappel, Carla Boetes, and Nico Karssemeijer. Computer-aided detection of masses at mammography: Interactive decision support versus prompts. *Radiology*, 266:123–129, 2013.

[128] Benjamin Q Huynh, Hui Li, and Maryellen L Giger. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *J Med Imaging*, 3:034501, Jul 2016.

[129] Sangheum Hwang and Hyo-Eun Kim. Self-transfer learning for fully weakly supervised object localization. *arXiv:1602.01625*, 2016.

[130] Debra Ikeda and Kanae Kawai Miyake. *Breast imaging: the requisites*. Elsevier Health Sciences, 2016.

[131] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. *arXiv preprint arXiv:150602025*, 2015.

[132] A. R. Jamieson, K. Drukker, and M. L. Giger. Breast image feature learning with adaptive deconvolutional networks. In *Medical Imaging*, volume 8315 of *Proceedings of the SPIE*, page 831506, 2012.

[133] Zhanglong Ji, Zachary C Lipton, and Charles Elkan. Differential privacy and machine learning: a survey and review. *arXiv:1412.7584*, 2014.

[134] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pages 675–678, 2014.

[135] Michiel Kallenberg, Kersten Petersen, Mads Nielsen, Andrew Ng, Pengfei Diao, Christian Igel, Celine Vachon, Katharina Holland, Nico Karssemeijer, and Martin Lillholm. Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. *IEEE Trans Med Imaging*, 35:1322–1331, 2016.

[136] 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. *Med Image Anal*, 36:61–78, 2017.

[137] Andrej Karpathy, George Toderici, Sachin Shetty, Tommy Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1725–1732. IEEE, 2014.

[138] N. Karssemeijer. Automated classification of parenchymal patterns in mammograms. *Phys Med Biol*, 43:365–378, 1998.

[139] N. Karssemeijer, J. D. Otten, A. A. J. Roelofs, S. van Woudenberg, and J. H. C. L. Hendriks. Effect of independent multiple reading of mammograms on detection performance. In *Medical Imaging*, volume 5372 of *Proceedings of the SPIE*, pages 82–89, 2004.

[140] N. Karssemeijer, J. D. M. Otten, H. Rijken, and R. Holland. Computer aided detection of masses in mammograms as decision support. *Br J Radiol*, 79 Spec No 2:S123–S126, 2006.

[141] Nico Karssemeijer. Personal communication, 2017.

[142] Nico Karssemeijer and Guido te Brake. Detection of stellate distortions in mammograms. *IEEE Trans Med Imaging*, 15:611–619, 1996.

[143] J. Kawahara, A. BenTaieb, and G. Hamarneh. Deep features to classify skin lesions. In *IEEE Int Symp Biomedical Imaging*, pages 1397–1400, 2016.

[144] Diederik Kingma and Jimmy Ba. ADAM: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

[145] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv:1312.6114*, 2013.

[146] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, page 201611835, 2017.

[147] Pavel Kisilev, Eli Sason, Ella Barkan, and Sharbell Hashoul. Medical image description using multi-task-loss CNN. In *International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, pages 121–129. Springer, 2016.

[148] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[149] T. Kooi, A. Gubern-Mérida, J. J. Mordang, R. Mann, R. Pijnappel, K. Schuur, A. den Heeten, and N. Karssemeijer. A comparison between a deep convolutional neural network and radiologists for classifying regions of interest in mammography. In A. Tingberg et al., editor, *Breast Imaging*, volume 9699 of *Lecture Notes in Computer Science*, pages 51–56. Springer International Publishing Switzerland, 2016.

[150] Thijs Kooi and Nico Karssemeijer. Invariant features for discriminating cysts from solid lesions in mammography. In *Breast Imaging*, pages 573–580. Springer, 2014.

[151] Thijs Kooi and Nico Karssemeijer. Classifying symmetrical differences and temporal change in mammography using deep neural networks. 2017.

[152] Thijs Kooi and Nico Karssemeijer. Deep learning of symmetrical discrepancies for computer-aided detection of mammographic masses. In *Proceedings of the SPIE*, 2017.

[153] 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. *Med Image Anal*, 35:303–312, 2016.

[154] Thijs Kooi, Bram van Ginneken, Nico Karssemeijer, and Ard den Heeten. Discriminating solitary cysts from soft tissue lesions in mammography using a pretrained deep convolutional neural network. *Medical physics*, 44:1017–1027, March 2017.

[155] Thijs Kooi, Bram van Ginneken, Nico Karssemeijer, and Ard den Heeten. Discriminating solitary cysts from soft tissue lesions in mammography using a pretrained deep convolutional neural network. *Medical Physics*, 2017.

[156] Alex. Krizhevsky, Ilya. Sutskever, and Geoffrey Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[157] Frank R Kschischang, Brendan J Frey, and Hans-Andrea Loeliger. Factor graphs and the sum-product algorithm. *Information Theory, IEEE Transactions on*, 47(2):498–519, 2001.

[158] Ashnil Kumar, Pradeeba Sridar, Ann Quinton, R Krishna Kumar, Dagan Feng, Ralph Nanan, and Jinman Kim. Plane identification in fetal ultrasound images using saliency maps and convolutional neural networks. In *IEEE Int Symp Biomedical Imaging*, pages 791–794, 2016.

[159] Sanjiv Kumar and Martial Hebert. Discriminative fields for modeling spatial dependencies in natural images. In *Advances in Neural Information Processing Systems*, 2003.

[160] M. A. Kupinski and M. L. Giger. Automated seeded lesion segmentation on digital mammograms. *IEEE Trans Med Imaging*, 17:510–517, 1998.

[161] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Int Conf Mach Learn*, pages 282–289, 2001.

[162] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278–2324, 1998.

[163] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[164] Constance D. Lehman, Robert D. Wellman, Diana S M. Buist, Karla Kerlikowske, Anna N A. Tosteson, Diana L. Miglioretti, and Breast Cancer Surveillance Consortium . Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern Med*, 175(11):1828–1837, Nov 2015.

[165] Guosheng Lin, Chunhua Shen, Anton van dan Hengel, and Reidm Ian. Efficient piecewise training of deep structured models for semantic segmentation. *arXiv:1504.01013*, 2015.

[166] Jianyu Lin, Neil T. Clancy, Xueqing Sun, Ji Qi, Mirek Janatka, Danail Stoyanov, and Daniel S. Elson. Probe-based rapid hybrid hyperspectral and tissue surface imaging aided by fully convolutional networks. In *Med Image Comput Comput Assist Interv*, volume 9902 of *Lect Notes Comput Sci*, pages 414–422, 2016.

[167] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv:1312.4400*, 2013.

[168] Zachary C Lipton. The mythos of model interpretability. *arXiv:1606.03490*, 2016.

[169] Zachary C Lipton, John Berkowitz, and Charles Elkan. A critical review of recurrent neural networks for sequence learning. *arXiv:1506.00019*, 2015.

[170] Zachary C Lipton, David C Kale, and Randall Wetzel. Modeling missing data in clinical time series with rnns. 2016.

[171] G. Litjens, O. Debats, J. Barentsz, N. Karssemeijer, and H. Huisman. Computer-aided detection of prostate cancer in MRI. *IEEE Trans Med Imaging*, 33:1083–1092, 2014.

[172] G. Litjens, T. Kooi, B. Ehteshami Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J.A.W.M. van der Laak, B. van Ginneken, and C.I. Sánchez. A survey on deep learning in medical image analysis. *arXiv:1702.05747*, 2017.

[173] Jiamin Liu, David Wang, Zhuoshi Wei, Le Lu, Lauren Kim, Evrim Turkbey, and Ronald M Summers. Colitis detection on computed tomography using regional convolutional neural networks. In *IEEE Int Symp Biomedical Imaging*, pages 863–866, 2016.

[174] Sheena Xin Liu. Symmetry and asymmetry analysis and its implications to computer-aided diagnosis: A review of the literature. 42(6):1056–1064, 2009.

[175] Yun Liu, Krishna Gadepalli, Mohammad Norouzi, George E Dahl, Timo Kohlberger, Aleksey Boyko, Subhashini Venugopalan, Aleksei Timofeev, Philip Q Nelson, Greg S Corrado, Jason D. Hipp, Lily Peng, and Martin C. Stumpe. Detecting cancer metastases on gigapixel pathology images. *arXiv:1703.02442*, 2017.

[176] S-CB Lo, S-LA Lou, Jyh-Shyan Lin, Matthew T Freedman, Minze V Chien, and Seong K Mun. Artificial convolution neural network techniques and applications for lung nodule detection. *IEEE Trans Med Imaging*, 14:711–718, 1995.

[177] Beth Logan et al. Mel frequency cepstral coefficients for music modeling. In *ISMIR*, 2000.

[178] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *arXiv:1411.4038*, 2015.

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

[180] Chris J Maddison, Aja Huang, Ilya Sutskever, and David Silver. Move evaluation in go using deep convolutional neural networks. *arXiv:14126564*, 2014.

[181] Martin A Makary and Michael Daniel. Medical error—the third leading cause of death in the us. *Bmj*, 353:i2139, 2016.

[182] Ansgar Malich, Dorothee R Fischer, and Joachim Böttcher. CAD for mammography: the technique, results, current role and further developments. *Eur Radiol*, 16:1449–1460, 2006.

[183] Andrew McAfee and Erik Brynjolfsson. The second machine age. *Work, Progress, and prosperity in time of brilliant technologies. New York: WW Norton & Company*, 2014.

[184] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. *arXiv:1606.04797*, 2016.

[185] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

[186] Pim Moeskops, Max A Viergever, Adrienne M Mendrik, Linda S de Vries, Manon J N L Benders, and Ivana Isgum. Automatic segmentation of MR brain images with a convolutional neural network. *IEEE Trans Med Imaging*, 35(5):1252–1262, 2016.

[187] Joris M. Mooij. libDAI: A free and open source C++ library for discrete approximate inference in graphical models. *J Mach Learn Res*, 11:2169–2173, Aug 2010.

[188] Mehdi Moradi, Yaniv Gur, Hongzhi Wang, Prasanth Prasanna, and Tanveer Syeda-Mahmood. A hybrid learning approach for semantic labeling of cardiac CT slices and recognition of body position. In *IEEE Int Symp Biomedical Imaging*, 2016.

[189] Hans Moravec. *Mind children*, volume 375. Cambridge Univ Press, 1988.

[190] J. J. Mordang, T. Janssen, A. Bria, T. Kooi, A. Gubern-Mérida, and N. Karssemeijer. Automatic microcalcification detection in multi-vendor mammography using convolutional neural networks. In *Breast Imaging*, volume 9699 of *Lect Notes Comput Sci*, pages 35–42, 2016.

[191] Jan-Jurre Mordang, Albert Gubern-Mérida, Gerard den Heeten, and Nico Karssemeijer. Reducing false positives of microcalcification detection systems by removal of breast arterial calcifications. *Med Phys*, 43(4):1676–1687, mar 2016.

[192] N. R. Mudigonda, R. M. Rangayyan, and J. E. Desautels. Gradient and texture analysis for the classification of mammographic masses. *IEEE Trans Med Imaging*, 19:1032–1043, 2000.

[193] Janne J. Nappi, Toru Hironaka, Daniele Regge, and Hiroyuki Yoshida. Deep transfer learning of virtual endoluminal views for the detection of polyps in CT colonography. In *Medical Imaging*, Proceedings of the SPIE, page 97852B, 2016.

[194] Natalia Neverova, Christian Wolf, Graham W Taylor, and Florian Nebout. Multi-scale deep learning for gesture detection and localization. In *European Conference on Computer Vision*, 2014.

[195] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. *arXiv:150308909*, 2015.

[196] M. Niemeijer, M. D. Abràmoff, and B. van Ginneken. Information fusion for diabetic retinopathy CAD in digital color fundus photographs. *IEEE Trans Med Imaging*, 28(5):775–785, 2009.

[197] Robert M Nishikawa. Current status and future directions of computer-aided diagnosis in mammography. *Comput Med Imaging Graph*, 31:224–235, 2007.

[198] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Comput Vis Pattern Recognit*, pages 1717–1724, 2014.

[199] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. Zero-shot learning with semantic output codes. In *Advances in neural information processing systems*, pages 1410–1418, 2009.

[200] Sridevi Parise and Max Welling. Learning in markov random fields: An empirical study. In *Joint Statistical Meeting*, 2005.

[201] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Attentive explanations: Justifying decisions and pointing to the evidence. *arXiv:1612.04757*, 2016.

[202] Chris Paxton, Alexandru Niculescu-Mizil, and Suchi Saria. Developing predictive models using electronic medical records: challenges and pitfalls. In *AMIA Annual Symposium Proceedings*, volume 2013, page 1109. American Medical Informatics Association, 2013.

[203] Judea Pearl. Reverend bayes on inference engines: A distributed hierarchical approach. In *Proceedings of the Second National Conference on Artificial Intelligence*, 1982.

[204] Markus Peura and Jukka Iivarinen. Efficiency of simple shape descriptors. In *Proceedings of the third international workshop on visual form*, volume 443, page 451, 1997.

[205] Ha Tran Hong Phan, Ashnil Kumar, Jinman Kim, and Dagan Feng. Transfer learning of a convolutional neural network for HEp-2 cell image classification. In *IEEE Int Symp Biomedical Imaging*, pages 1208–1211, 2016.

[206] John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. 10(3):61–74, 1999.

[207] Adhish Prasoon, Kersten Petersen, Christian Igel, François Lauze, Erik Dam, and Mads Nielsen. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. In *Med Image Comput Comput Assist Interv*, volume 8150 of *Lect Notes Comput Sci*, pages 246–253, 2013.

[208] Adhish Prasoon, Kersten Petersen, Christian Igel, François Lauze, Erik Dam, and Mads Nielsen. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. *Med Image Comput Comput Assist Interv*, 16:246–253, 2013.

[209] Yuchen Qiu, Yunzhi Wang, Shiju Yan, Maxine Tan, Samuel Cheng, Hong Liu, and Bin Zheng. An initial investigation on developing a new method to predict short-term breast cancer risk based on deep learning technology. In *Medical Imaging*, volume 9785 of *Proceedings of the SPIE*, page 978521, 2016.

[210] Ariadna Quattoni, Michael Collins, and Trevor Darrell. Conditional random fields for object recognition. In *Advances in neural information processing systems*, pages 1097–1104, 2004.

[211] Martin Rajchl, Matthew CH Lee, Ozan Oktay, Konstantinos Kamnitsas, Jonathan Passerat-Palmbach, Wenjia Bai, Bernhard Kainz, and Daniel Rueckert. DeepCut: Object segmentation from bounding box annotations using convolutional neural networks. *IEEE Trans Med Imaging*, 2016. , in press.

[212] R. M. Rangayyan, N. M. El-Faramawy, J. E. L. Desautels, and O. A. Alim. Measures of acutance and shape for classification of breast tumors. *IEEE Trans Med Imaging*, 16:799–810, 1997.

[213] Vijay M Rao, David C Levin, Laurence Parker, Barbara Cavanaugh, Andrea J Frangos, and Jonathan H Sunshine. How widely is computer-aided detection used in screening and diagnostic mammography? *J Am Coll Radiol*, 7:802–805, 2010.

[214] Daniele Ravi, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, and Guang-Zhong Yang. Deep learning for health informatics. *IEEE J Biomed Health Inform*, 21:4–21, January 2017.

[215] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: An astounding baseline for recognition. *arXiv:1403.6382*, 2014.

[216] A. A. J. Roelofs, N. Karssemeijer, N. Wedekind, C. Beck, S. van Woudenberg, P. R. Snoeren, J. H. C. L. Hendriks, M. Rosselli del Turco, N. Bjurstam, H. Junkermann, D. Beijerinck, B. Séradour, and C. J. G. Evertsz. Importance of comparison of current and prior mammograms in breast cancer screening. *Radiology*, 242:70–77, 2007.

[217] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Med Image Comput Comput Assist Interv*, volume 9351 of *Lect Notes Comput Sci*, pages 234–241, 2015.

[218] H. R. Roth, L. Lu, J. Liu, J. Yao, A. Seff, K. Cherry, L. Kim, and R. M. Summers. Improving computer-aided detection using convolutional neural networks and random view aggregation. *IEEE Trans Med Imaging*, 35(5):1170–1181, 2016.

[219] H. R. Roth, L. Lu, A. Seff, K. M. Cherry, J. Hoffman, S. Wang, J. Liu, E. Turkbey, and R. M. Summers. A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations. In *Med Image Comput Comput Assist Interv*, volume 8673 of *Lect Notes Comput Sci*, pages 520–527, 2014.

[220] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet large scale visual recognition challenge. *Int J Comput Vis*, 115(3):1–42, 2014.

[221] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, 2 edition, 2003.

[222] Sepideh Saadatmand, Reini Bretveld, Sabine Siesling, and Madeleine MA Tilanus-Linthorst. Influence of tumour stage at breast cancer detection on survival in modern times: population based study in 173 797 patients. *British Medical Journal Publishing Group*, 2015.

[223] B. Sahiner, Heang-Ping Chan, N. Petrick, Datong Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsitt. Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images. *IEEE Trans Med Imaging*, 15:598–610, 1996.

[224] Ravi K. Samala, Heang-Ping Chan, Lubomir Hadjiiski, Kenny Cha, and Mark A. Helvie. Deep-learning convolution neural network for computer-aided detection of microcalcifications in digital breast tomosynthesis. In *Medical Imaging*, volume 9785 of *Proceedings of the SPIE*, page 97850Y, 2016.

[225] Ravi K Samala, Heang-Ping Chan, Lubomir Hadjiiski, Mark A Helvie, Jun Wei, and Kenny Cha. Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography. *Medical Physics*, 43(12):6654–6666, 2016.

[226] M. Samulski, R. Hupse, C. Boetes, R. Mus, G. den Heeten, and N. Karssemeijer. Using Computer Aided Detection in Mammography as a Decision Support. *Eur Radiol*, 20:2323–2330, 2010.

[227] M. Samulski and N. Karssemeijer. Optimizing Case-based Detection Performance in a Multiview CAD System for Mammography. *IEEE Trans Med Imaging*, 30:1001–1009, 2011.

[228] Suchi Saria, Anand K Rajani, Jeffrey Gould, Daphne Koller, and Anna A Penn. Integration of early physiological responses predicts later illness severity in preterm infants. *Science translational medicine*, 2(48):48ra65–48ra65, 2010.

[229] Jürgen Schmidhuber. Deep learning in neural networks: an overview. *Neural Netw*, 61:85–117, 2015.

[230] J. Schwaab, A. Gubern-Mérida, L. Wang, and M. Gunther. Automatic assessment of nipple position in Automated 3D Breast Ultrasound images. In *MICCAI Workshop: Breast Image Analysis*, 2015.

[231] Alexander G Schwing and Raquel Urtasun. Fully connected deep structured networks. *arXiv:1503.02351*, 2015.

[232] Diane Scutt, Gillian A Lancaster, and John T Manning. Breast asymmetry and predisposition to breast cancer. 8(2):1, 2006.

[233] A. A. A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S. van Riel, M. Winkler Wille, M. Naqibullah, C. Sanchez, and B. van Ginneken. Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks. *IEEE Trans Med Imaging*, 35(5):1160–1169, 2016.

[234] Hoo-Chang Shin, Le Lu, Lauren Kim, Ari Seff, Jianhua Yao, and Ronald M Summers. Interleaved text/image deep mining on a very large-scale radiology database. In *Comput Vis Pattern Recognit*, pages 1090–1099, 2015.

[235] Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M Summers. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. *arXiv:1603.08486*, 2016.

[236] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans Med Imaging*, 35(5):1285–1298, 2016.

[237] Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1310–1321. ACM, 2015.

[238] Edward H Shortliffe. Mycin: A knowledge-based computer program applied to infectious diseases. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 66. American Medical Informatics Association, 1977.

[239] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *Int J Comput Vis*, 81, 2009.

[240] Edward A Sickles. Probably benign breast lesions: When should follow-up be recommended and what is the optimal follow-up protocol? *Radiology*, 213(1):11–14, 1999.

[241] Edward A Sickles. The spectrum of breast asymmetries: imaging features, work-up, management. 45(5):765–771, 2007.

[242] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2016. *CA Cancer J Clin*, 66(1):7–30, 2016.

[243] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529:484–489, 2016.

[244] P. Y. Simard, D. Steinkraus, and J. C. Platt. Best practices for convolutional neural networks applied to visual document analysis. In *Document Analysis and Recognition*, pages 958–963, 2003.

[245] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, 2014.

[246] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.

[247] Per Skaane, Andriy I. Bandos, Randi Gullien, Ellen B. Eben, Ulrika Ekseth, Unni Haakenaasen, Mina Izadi, Ingvild N. Jebsen, Gunnar Jahr, Mona Krager, Loren T. Niklason, Solveig Hofvind, and David Gur. Comparison of digital mammography alone and digital mammography plus tomosynthesis in a population-based screening program. *Radiology*, 267:47–56, 2013.

[248] Per Skaane, Randi Gullien, Hilde Bjørndal, Ellen B. Eben, Ulrika Ekseth, Unni Haakenaasen, Gunnar Jahr, Ingvild Naess Jebsen, and Mona Krager. Digital breast tomosynthesis (dbt): initial experience in a clinical setting. *Acta Radiol*, 53:524–529, 2012.

[249] Per Skaane, Ashwini Kshirsagar, Sandra Stapleton, Kari Young, and Ronald A Castellino. Effect of computer-aided detection on independent double reading of paired screen-film and full-field digital screening mammograms. pages 377–384, 2007.

[250] P. R. Snoeren and N. Karssemeijer. Thickness correction of mammographic images by means of a global parameter model of the compressed breast. *IEEE Trans Med Imaging*, 23:799–806, 2004.

[251] Youyi Song, Ling Zhang, Siping Chen, Dong Ni, Baiying Lei, and Tianfu Wang. Accurate segmentation of cervical cytoplasm and nuclei based on multiscale convolutional network and graph partitioning. *IEEE Trans Biomed Eng*, 62(10):2421–2433, 2015.

[252] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J Mach Learn Res*, 15(1):1929–1958, 2014.

[253] Marijn F Stollenga, Wonmin Byeon, Marcus Liwicki, and Juergen Schmidhuber. Parallel multi-dimensional LSTM, with application to fast biomedical volumetric image segmentation. In *Advances in Neural Information Processing Systems*, pages 2998–3006, 2015.

[254] Natasha K. Stout, Sandra J. Lee, Clyde B. Schechter, Karla Kerlikowske, Oguzhan Alagoz, Donald Berry, Diana S M. Buist, Mucahit Cevik, Gary Chisholm, Harry J. de Koning, Hui Huang, Rebecca A. Hubbard, Diana L. Miglioretti, Mark F. Munsell, Amy Trentham-Dietz, Nicolien T. van Ravesteyn, Anna N A. Tosteson, and Jeanne S. Mandelblatt. Benefits, harms, and costs for breast cancer screening after us implementation of digital mammography. *J Natl Cancer Inst*, 106:dju092, 2014.

[255] Masashi Sugiyama and Motoaki Kawanabe. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press, 2012.

[256] Wenqing Sun, Tzu-Liang Bill Tseng, Jianying Zhang, and Wei Qian. Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data. *Comput Med Imaging Graph*, 2016.

[257] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *Int Conf Mach Learn*, pages 1139–1147, 2013.

[258] Charles Sutton and Andrew McCallum. An introduction to conditional random fields. *Machine Learning*, 2011.

[259] Charles Sutton and Andrew McCallum. Piecewise training for undirected models. *arXiv:1207.1409*, 2012.

[260] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *arXiv:1409.4842*, 2014.

[261] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv:1312.6199*, 2013.

[262] Laszlo Tabar, Ming-Fang Yen, Bedrich Vitak, Hsiu-Hsi Tony Chen, Robert A Smith, and Stephen W Duffy. Mammography service screening and mortality in breast cancer patients: 20-year follow-up before and after introduction of screening. *The Lancet*, 361:1405–1410, 2003.

[263] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Fine tuning or full training? *IEEE Trans Med Imaging*, 35(5):1299–1312, 2016.

[264] P. Taylor, J. Champness, R. Given-Wilson, K. Johnston, and H. Potts. Impact of computer-aided detection prompts on the sensitivity and specificity of screening mammography. *Health Technol Assess*, 9:iii, 1–iii,58, 2005.

[265] G. M. te Brake and N. Karssemeijer. Segmentation of suspicious densities in digital mammograms. *Med Phys*, 28:259–266, 2001.

[266] G. M. te Brake, N. Karssemeijer, and J. H. Hendriks. An automatic method to discriminate malignant masses from normal tissue in digital mammograms. *Phys Med Biol*, 45:2843–2857, 2000.

[267] M Gelig Thurfjell, B Vitak, E Azavedo, G Svane, and E Thurfjell. Effect on sensitivity and specificity of mammography screening with or without comparison of old mammograms. *Acta Radiologica*, 41(1):52–56, 2000.

[268] S. Timp and N. Karssemeijer. Interval change analysis to improve computer aided detection in mammography. *Med Image Anal*, 10:82–95, 2006.

[269] S. Timp, C. Varela, and N. Karssemeijer. Temporal change analysis for characterization of mass lesions in mammography. *IEEE Trans Med Imaging*, 26:945–953, 2007.

[270] Sheila Timp and Nico Karssemeijer. A new 2D segmentation method based on dynamic programming applied to computer aided detection in mammography. *Medical Physics*, 31(5):958–971, 2004.

[271] Sheila Timp and Nico Karssemeijer. A new 2D segmentation method based on dynamic programming applied to computer aided detection in mammography. *Med Phys*, 31:958–971, 2004.

[272] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. In *Utility, probability, and human decision making*, pages 141–162. Springer, 1975.

[273] Jasper RR Uijlings, Koen EA van de Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *Int J Comput Vis*, 104:154–171, 2013.

[274] S. van Engeland, P. Snoeren, J. Hendriks, and N. Karssemeijer. A comparison of methods for mammogram registration. *IEEE Trans Med Imaging*, 22:1436–1444, 2003.

[275] Saskia van Engeland, Peter Snoeren, Henk-Jan Huisman, Carla Boetes, and Nico Karssemeijer. Volumetric breast density estimation from full-field digital mammograms. *IEEE Transactions on Medical Imaging*, 25(3):273–282, March 2006.

[276] B. van Ginneken. *Computer-aided diagnosis in chest radiography*. PhD thesis, Utrecht University, The Netherlands, 2001.

[277] B. van Ginneken, C. M. Schaefer-Prokop, and M. Prokop. Computer-aided diagnosis: how to move from the laboratory to the clinic. *Radiology*, 261(3):719–732, 2011.

[278] Bram van Ginneken, Arnaud AA Setio, Colin Jacobs, and Francesco Ciompi. Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. In *IEEE Int Symp Biomedical Imaging*, pages 286–289, 2015.

[279] Mark J. J. P. van Grinsven, Bram van Ginneken, Carel B. Hoyng, Thomas Theelen, and Clara I. Sánchez. Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images. *IEEE Trans Med Imaging*, 35(5):1273–1284, 2016.

[280] C. Varela, N. Karssemeijer, J. H. C. L. Hendriks, and R. Holland. Use of prior mammograms in the classification of benign and malignant masses. *Eur J Radiol*, 56:248–255, 2005.

[281] Wouter J. H. Veldkamp and Nico Karssemeijer. Improved method for detection of microcalcification clusters in digital mammograms. In *Medical Imaging*, volume 3661 of *Proceedings of the SPIE*, pages 512–522, 1999.

[282] M. Velikova, M. Samulski, P. J. F. Lucas, and N. Karssemeijer. Improved mammographic CAD performance using multi-view information: a Bayesian network framework. *Phys Med Biol*, 54:1131–1147, 2009.

[283] Marina Velikova, Peter J F. Lucas, Maurice Samulski, and Nico Karssemeijer. A probabilistic framework for image information fusion with an application to mammographic analysis. *Med Image Anal*, 16:865–875, 2012.

[284] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J Mach Learn Res*, 11:3371–3408, 2010.

[285] Chaohui Wang, Nikos Komodakis, and Nikos Paragios. Markov random field modeling, inference & learning in computer vision & image understanding: A survey. *Comput Vis Image Underst*, 117(11):1610–1627, 2013.

[286] Juan Wang, Huanjun Ding, FateMeh Azamian, Brian Zhou, Carlos Iribarren, Sabee Molloi, and Pierre Baldi. Detecting cardiovascular disease from mammograms with deep learning. *IEEE Trans Med Imaging*, 2017.

[287] Max Welling and Charles A Sutton. Learning in markov random fields with contrastive free energies. In *AISTATS*, 2005.

[288] Janine Willis and Alexander Todorov. First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological science*, 17(7):592–598, 2006.

[289] Wei Yang, Yingyin Chen, Yunbi Liu, Liming Zhong, Genggeng Qin, Zhentai Lu, Qianjin Feng, and Wufan Chen. Cascade of multi-scale convolutional neural networks for bone suppression of chest radiographs in gradient domain. *Med Image Anal*, 35:421–433, 2016.

[290] B. C. Yankaskas, M. J. Schell, R. E. Bird, and D. A. Desrochers. Reassessment of breast cancers missed during routine screening mammography: a community-based study. *AJR Am J Roentgenol*, 177:535–541, 2001.

[291] Bonnie C Yankaskas, Ryan C May, Jeanine Matuszewski, J Michael Bowling, Molly P Jarman, and Bruce F Schroeder. Effect of observing change from comparison mammograms on performance of screening mammography in a large community-based population. *Radiology*, 261(3):762–770, 2011.

[292] Jonathan S Yedidia, William T Freeman, and Yair Weiss. Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, 8:236–239, 2003.

[293] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, pages 3320–3328, 2014.

[294] Ji Hyun Youk, Eun-Kyung Kim, Kyung Hee Ko, and Min Jung Kim. Asymmetric mammographic findings based on the fourth edition of bi-rads: Types, evaluation, and management 1. 29(1):e33–e33, 2009.

[295] Sophia Zackrisson, Ingvar Andersson, Lars Janzon, Jonas Manjer, and Jens Peter Garne. Rate of over-diagnosis of breast cancer 15 years after end of malmö mammographic screening trial: follow-up study. *Bmj*, 332(7543):689–692, 2006.

[296] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002.

[297] Qi Zhang, Yang Xiao, Wei Dai, Jingfeng Suo, Congzhi Wang, Jun Shi, and Hairong Zheng. Deep learning based classification of breast tumors with shear-wave elastography. *Ultrasonics*, 72:150–157, 2016.

[298] Bin Zheng, Xingwei Wang, Dror Lederman, Jun Tan, and David Gur. Computer-aided detection; the effect of training databases on detection of subtle breast masses. *Acad Radiol*, 17:1401–1408, 2010.

[299] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip Torr. Conditional random fields as recurrent neural networks. In *Int Conf Comp Vis*, 2015. arXiv:1502.03240.

Acknowledgements

This research was funded by grant KUN 2012-5577 of the Dutch Cancer Society and supported by the Foundation of Population Screening Mid West.