Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Monterrey

School of Engineering and Information Technology

Graduate Program

Master of Science in Intelligent Systems

Thesis Proposal

Early Detection and Diagnosis of Breast Cancer Lesions in Digital Mammograms using Deep Convolutional Networks

by

Erick Michael Cobos Tandazo
1184587

Monterrey, N.L., May 14, 2015

Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Monterrey

School of Engineering and Information Technology

Graduate Program

The committee members, hereby, recommend that the master's thesis proposal presented by Erick Michael Cobos Tandazo be accepted to develop the thesis project as a partial requirement for the degree of **Master of Science**, with a major in:

Intelligent Systems

Thesis Committee:

Dr. Hugo Terashima Marín
Principal Advisor

Por definir
Committee Member

Por definir
Committee Member

Dr. Ramón Brena Pinero
Director of the Master's Program in
Intelligent Systems

May 14, 2015

Contents

1 Introduction **1**

2 Problem **2**

3 Objectives **4**

4 Hypothesis **4**
4.1 Research Questions 5

5 Background **5**
5.1 Breast Cancer 5
5.1.1 Mammograms 6
5.2 Classification 8
5.3 Artificial Neural Networks 11
5.4 Convolutional Networks 13
5.5 Convolutional Networks applied to Breast Cancer 13
5.5.1 Related Work 13

6 Methodology **14**

7 Work Plan **15**

Abstract

Breast cancer is one of the most common and deadliest cancer in woman around the world. The best tools used today for early breast cancer diagnosis are screening mammograms; mammograms are x-ray pictures of the breast used by radiologists to identify microcalcifications and breast masses, signs of early breast cancer development. Traditional computer systems use handmade features and complex image techniques to detect these lesions in mammographic images. In this work, we plan to use convolutional networks, a recent development in computer vision, which can automatically learn the relevant features for the classification task given enough training data. Convolutional networks have been used in some studies for breast cancer detection but we hope to introduce newer features and carefully tune the architecture to produce improved results. Additionally, this will be the first approximation to use deep learning techniques as part of an ongoing project in the institution which aims to develop a computer-aided diagnosis system for breast cancer. This thesis proposal is presented for approval to obtain the degree of Master of Science in Intelligent Systems.

1 Introduction

Automatic breast cancer diagnosis is a very difficult task for current computational systems. In this thesis, we apply deep learning techniques to digital mammographic images in order to improve the performance of such systems. We layout here the hypotheses, experiments and goals of our future research.

Breast cancer is a disease caused by abnormal breast cells which grow out of control forming tumors and invading surrounding tissue. It has the highest incidence rate of any cancer in the United States, an estimated 14.1% of all new cancer cases in 2015 will be breast cancer, and the third highest mortality of any cancer accounting for 6.9% of cancer related deaths. Among women it is by large the most commonly diagnosed cancer (28.6% of all cancers) and has the highest death rate (14.5%) besides lung cancer [American Cancer Society, 2015].

The recommended method for early breast cancer detection in aging women is to have regular screening mammograms. Mammograms are x-ray images of the breast used by radiologists to look for signs of possible tumor formation. There are different lesions that can be found on a mammogram, we focus on two: clustered microcalcifications, tiny deposits of calcium which could appear around cancerous tissue, and breast masses, more direct signs of the existence of a tumor although they are very often benign. Most breast cancers can be detected with a mammogram.

In this work, we center on using mammograms to automatically detect microcalcifications and breast masses and predict the probability of breast cancer on the patient. Although manual examination of mammograms has a high sensitivity rate, automatic analysis could be used on places where expert radiologists are not available or it could be used by doctors as a second informed opinion or to help them decide to which regions of the image dedicate more time. With this motivation, a project that intends to design a computer aided diagnosis (CAD) tool for breast cancer has existed in this institution since 2007. This thesis falls under the scope of this project and will be its first approximation to use deep learning for breast cancer diagnosis.

Traditional CAD systems for breast cancer diagnosis work as a pipeline where each stage uses different computer vision and machine learning techniques. An standard pipeline will, for instance, preprocess the image, identify and segment the relevant parts of the picture, extract features from the segmented parts and train a classifier on the extracted features. Although

some successful systems are built in this manner, they have a few disadvantages: each stage is a separate component and hence work is needed on each of them to notably improve overall results, it is composed of dependent stages so that changes on one component can affect the performance of other parts of the system, it uses complex image vision techniques to segment the images and extract features which are difficult to handcraft and select, it requires expert knowledge to be properly tuned, among others.

We plan to investigate the potential of convolutional networks to replace some if not all of the stages of traditional image processing systems. Convolutional networks [Fukushima, 1980, LeCun et al., 1998], a natural extension to feedforward neural networks, are a statistical learning classifier which uses raw images as input and learns the important features for the classification task as it is trained. Convolutional networks are designed to work with minimally preprocessed images, can be trained to be rotational and translational invariant and perform segementation, feature extraction and classification in one step. In our case, convolutional networks simplify the process of classification potentially reducing it to one component which is trained from labelled data and can be improved and properly tuned to obtain better results. Although there are some drawbacks with convolutional networks, they are the state-of-the-art technology for object recognition [Russakovsky et al., 2014] and we believe it is worthy to experiment with them.

We will start our experiments training a simple convolutional network in images preprocessed with different techniques including unpreprocessed images, later we will train a convolutional network using whole mammogram images, pretrain a convolutional network with a different image database and fine-tune it using our database and finally we will use the gathered knowledge to build an optimal convolutional network. We intend to learn whether convolutional networks can automatically preprocess mammographic images or else which is the best preprocessing for mammographic images, what is the best segmentation strategy we can use and whether we can achieve results similar to those of more traditional systems.

This document starts by offering an insight into the problem with traditional methods for image analysis in Section 2. It exposes the particular objectives and hypotheses of the thesis in Sections 3 and 4. Section 5 presents a comprehensive background of the scientific concepts used throughout the document and lastly a detailed methodology and work plan are shown in Sections 6 and 7.

2 Problem

Breast cancer is the most commonly diagnosed cancer in woman and its death rates are among the highest of any cancer. It is estimated that about 1 in 8 U.S. women will be diagnosed with breast cancer at some point in their lifetime. Early detection is key in reducing the number of deaths from breast cancer; detection in its earlier stage (*in situ*) increases the survival rate to virtually 100% [Howlader et al., 2014].

With current technology, a high quality mammogram is "the most efective way to detect breast cancer early" [National Cancer Institute, 2014]. Mammograms are x-ray images of each breast used by radiologists to search for early signs of cancer such as tumors or microcalcifications. About 85% of breast cancers can be detected with a screening mammogram [Breast Cancer Surveillance Consortium, 2013]. This high sensitivity is the product of the careful examination of the mammograms by experienced radiologists. A computer-aided diagnosis tool (CAD) could automatically detect and diagnose these abnormalities saving the

time and training needed by expert radiologists and avoiding any human error. Computer based approaches could also be used by radiologists as a help during the screening proccess or as a second informed opinion on a diagnostic.

CAD systems are based on image and classification techniques coming from Artificial Intelligence and Machine Learning. Traditional CAD tools for breast cancer diagnosis are composed of three steps: feature extraction, feature selection and classification. In the feature extraction phase, the system uses filters and image transformations to preprocess the mammogram and find geometric patterns which are used to produce a set of features for the image; expert knowledge is sometimes used in this phase. Feature selection or regularization is used to focus only on the important features for the classification task. Once a vector of features is obtained for each image, an standard binary classifier can be used to perform the final detection or diagnosis. These techniques have been used for many years and are standard in the industry [1].

Despite its widespread use and efficiency, systems based on traditional computer vision techniques have various limitations that should be addressed to further improve its performance:

- There is no standard way of preprocessing mammograms. Some filters are commonly used but their performances can vary.
- It uses handcrafted features. The features extracted from the image are chosen beforehand (maybe designed with the help of experts) and special filters and image techniques are used to extract them.
- It normally uses a small patch of the mammogram and makes a prediction on that patch but it does not consider the entire mammogram neither to make a prediction on the patient or to account for correlation between patches.
- To produce good results it requires knowledge in various fields such as radiology, oncology, image processing, computer vision, machine learning, etc.
- It is composed of many sequential steps. At each stage, there are many techniques from which the researcher can choose and many parameters which have to be estimated. This represents a cost in time and results as it is improbable that the optimal selection of techniques and parameters is achieved.
- As it is a complex system with different subsystems involved many other issues can arise such as non desired or unknown dependencies between subsystems, difficulty to localize errors, maintainability, etc.
- The techniques currently used are complex but the improvements achieved are not substantial. Much work is needed to make only incremental improvements and it is hard to know to which part of the system dedicate more resources.

This project will center around using Convolutional Networks, a recent development in computer vision, (see Section 5.4) to tackle some of these limitations, especifically automate preprocessing and feature extraction, use entire mammogram images and simplify the system pipeline by using a convolutional network as a replacement for many steps traditionally performed in succesion.

[1] See [Hernandez, 2014] for an example of a CAD system developed in this institution.

3 Objectives

The main goal of this work is to succesfully apply convolutional networks in digital mammograms to detect and diagnose breast cancer lessions, microcalcifications and breast masses, and to compare our results to those obtained by other groups working in convolutional networks for breast cancer diagnosis.

Particularly, there are various subgoals which we expect to achieve as the project advances:

- Develop a working pipeline for processing the mammographic images from our database and training a convolutional network. Essentially, this tool could also be used for other image classification tasks.
- Use a simple convolutional network to perform detection and diagnosis and study these initial results to guide further research.
- Show the viability of convolutional networks for breast cancer diagnosis.
- Use convolutional networks on an entire mammogram instead of only on small patches.
- Analyze the performance of convolutional networks reported on the literature.
- Use the improved convolutional network in the IRMA database.
- Generate results that could produce a conference or journal article.
- Propose new ideas and methods for future research in the topic.

Initial exploratory research has not yet been performed and some of these particular objectives may be modified as the project progresses. Furthermore, some new research avenues could be taken if they seem promising, for instance, using convolutional networks with digital tomosynthesis images (3-dimensional x-ray images of the breast).

4 Hypothesis

Although a considerable amount of work on breast cancer detection and diagnosis has been done in the institution, this project will be the first approximation to using convolutional networks for efficiently detecting and diagnosing breast cancer. Convolutional networks are widely used for object recognition tasks and have shown very good results [Russakovsky et al., 2014, Taigman et al., 2014, Dieleman et al., 2015]. They have a big research community and have become one of the preferred methods to perform image classification tasks.

Due to the exploratory nature of this work we are not truly certain of the results that will be obtained. Nevertheless, we have a well established idea of what we expect to obtain. We will apply some of these newly developed techniques expecting to produce similar or better results than those obtained using more traditional computer vision techniques. We believe that implementing convolutional networks for mamographic images will not be very difficult as it has already been done (see Section 5.5). We do not think that a simple convolutional network will suffice to obtain acceptable results; we will need to use a more refined convolutional network with well fitted parameters.

4.1 Research Questions

Some of the questions which will be answered in this work are:

- Can we improve the results reported by other groups using convolutional networks? Is training a convolutional network on mammographic images better than computing numeric features from the mammograms and training a simple classifier?
- Is deep learning feasible with the resources we have? Is our data and computational power sufficient? Is there any advantage to use GPU acceleration?
- Can we simplify the pipeline for breast cancer diagnosis? Can preprocessing be replaced by more layers on the same convolutional network? Could we use an entire mammogram for diagnosis instead of only small patches or could we automatically join results for small patches to generate results on the entire mammogram?
- What are the best parameters for our convolutional networks (number of layers, number of units, kernel sizes, regularization, activation functions, etc)? Is there a big improvement on refining the network and tuning parameters?
- What are the advantages of using a deep versus a shallow convolutional network?
- Could we use a convolutional network trained on a different database (such as the ImageNet database) to obtain features for mammographic images and use these features for classification?
- Are convolutional networks a good option for future research?

5 Background

We offer an introduction to some of the essential concepts needed to understand the rest of this document. We start by discussing breast cancer and mammograms in Section 5.1, we offer some basic concepts about classification and evaluation metrics in Section 5.2, in Sections 5.3 and 5.4 we give a short introduction into Artificial Neural Networks and Convolutional Neural Networks and finally we offer an overview of how convolutional networks have been used for breast cancer diagnosis in Section 5.5.

5.1 Breast Cancer

Cancer is an umbrella term to refer to a group of diseases caused by abnormal cell growth in different parts of the body. The accumulation of extra cells usually forms a mass of tissue called a *tumor*. Tumors can be benign or malignant: *benign tumors* are noncancerous, lack the ability to invade surrounding tissue and will not regrow if removed from the body; malignant or *cancerous tumors* are harmful, can invade nearby organs and tissues (*invasive cancer*), can spread to other parts of the body (*metastasis*) and will sometimes regrow when removed [National Cancer Institute, 2012].

Breast cancer is the cancer that forms in tissues of the breast. The two most common types of breast cancer are *ductal carcinoma* and *lobular carcinoma*; these cancers begin in the breast ducts and lobules, respectively (see Fig. 1). Breast cancer *incidence rate*, the number of new cases in a specified population during a year, is the highest of any cancer among

American women. Its *mortality rate*, the number of deaths during a year, is also one of the highest of any cancer [Howlader et al., 2014].

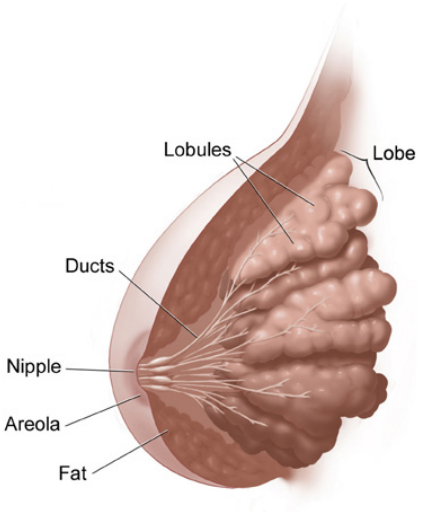

Figure 1: Anatomy of the female breast. Image courtesy of NCI.

The *cancer stage* depends on the size of the tumor and whether the cancer cells have spread to neighboring tissue or other parts of the body. It is expressed as a Roman numeral ranging from 0 through IV; stage I cancer is considered *early-stage breast cancer* and breast cancer at stage IV is considered *advanced*. Stage 0 describes non-invasive breast cancers, also known as *carcinoma in situ*. Stage I, II and III describe invasive breast cancer, i.e., cancer that has invaded normal surrounding breast tissue. Stage IV is used to describe metastatic cancer, i.e., breast cancer has spread beyond nearby tissue to other organs of the body.

5.1.1 Mammograms

A *mammogram* is an x-ray image of the breast. *Screening mammograms* (normally composed of two mammograms of each breast) are used to check for breast cancer signs on women who have not shown symptoms of the disease. If an abnormality is found, a *diagnostic mammogram* is ordered, these are detailed x-ray pictures of the suspicious region [National Cancer Institute, 2014]. A standard mammogram is shown in Fig. 2.

Having a screening mammogram in a regular basis is the most effective method for detecting early breast cancer; around 85% of breast cancers can be detected in a screening mammogram [Breast Cancer Surveillance Consortium, 2013]. Nevertheless, screening mammograms have many limitations: a high false positive rate, overtreatment in Stage 0 cancer, false negative results for women with high breast density, radiation exposure and physical and psychological discomfort [National Cancer Institute, 2014].

Mammograms are read by expert radiologists. The radiologist looks primarily for microcalcifications and breast masses. *Microcalcifications* are tiny deposits of calcium in the breast tissue which can be a sign of early breast cancer if found in clusters with irregular layout and shapes. *Breast masses* or breast lumps are possibly a variety of things: fluid-filled cysts, fatty tissues, fibric tissues, noncancerous or cancerous tumors, among others. A mass can be a sign of breast cancer if it has poorly defined shape and margins. See Fig. 3 for an example of possible signs of breast cancer. Radiologists will also consider the breast density of the patient when

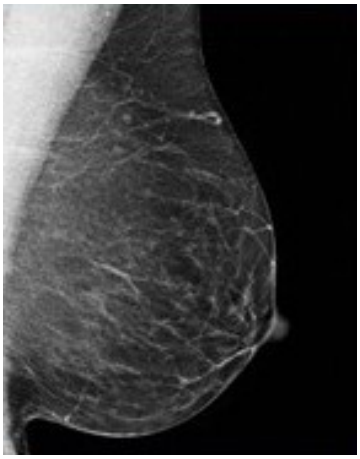

Figure 2: A standard mammogram.

reading a mammogram given that high breast density is linked to a higher risk of breast cancer and it also difficults the interpretation of the mammogram [American Cancer Society, 2014].

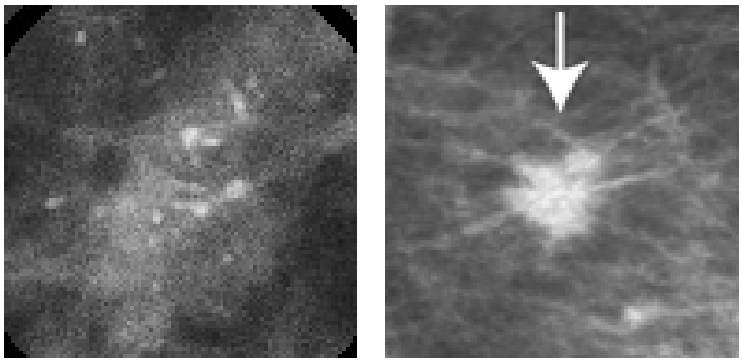

Figure 3: Signs of possible breast cancer in a mammogram. Left: A cluster of microcalcifications in an irregular layout. Right: A poorly defined breast mass.

Conventional mammography uses film to record x-ray images of the breast. *Digital mammography* on the other hand uses digital receptors to convert the x-rays into electric signals and stores the image electronically. Digital mammograms offer a clearer picture of the breast and can be digitally manipulated and shared between health care providers. Its effectiveness to identify breast cancer over film mammograms, however, is still debated [Kerlikowske et al., 2011, Pisano et al., 2008, Skaane et al., 2007]. Digital mammography is steadily becoming the standard for breast cancer screening, Fig. 2 is, in fact, a digital mammogram.

Digital tomosynthesis, also called three-dimensional mammography, is a new technology that essentially produces 3-dimensional x-ray images of the breast and is expected to improve the efficacy of regular 2-d mammograms. Studies comparing the two techniques have not yet been published, though [National Cancer Institute, 2014].

In this thesis we will center on using mammograms, either digital or manually digitized from film, to detect microcalcifications and masses and predict the likelihood of breast cancer on the patient.

Much of this section was written using information from the National Cancer Institute. We recommend to visit its website (`www.cancer.gov`) for more information.

5.2 Classification

Machine learning is the study of algorithms that build models of a population or of a function of interest and estimate their parameters from data in order to make predictions or inferences. A machine learning expert knows how to choose the right model for the problem in hand (*model selection*), how to efficiently estimate its parameters from the available data (*learning* or *training phase*) and how to evaluate the trained model (*testing phase*).

Machine learning problems can be divided into three categories depending on the data used to train the model: *supervised learning*, where we learn a function $f(x)$ using a set of examples which are labelled with the correct output, for instance, learning a function that estimates the price of a house given its size and number of bedrooms from a dataset of houses labelled with their real value; *unsupervised learning*, where we look for relationships and structure in unlabelled data, for instance, given a dataset of potential customers find those who are likely to buy a car and *reinforcement learning*, where the only feedback received are rewards, for example, learning to play tetris from a dataset of world states and actions and where rewards are received sparsely every time points are earned (when lines dissapear). Supervised learning can be further divided in regression and classification. If the expected output is numerical, e.g., the price of a house, it is called *regression*, if the expected ouput is categorical, e.g., spam or no spam, it is called *classification*. We will focus on classification.

A *classifier* takes as input a vector of *features* $x \in \mathbb{R}^n$ from a problem instance and produces an *output* $h(x)$ predicting the class y that instance belongs to, i.e., it concretely models the underlying function $f(x)$ as $h(x)$ (h stands for hypothesis). *Binary classification*, when y can only take two values e.g., cancer/no cancer, is the most common kind of classification and *multiclass classification*, when y can take $K > 2$ different values, can be performed by using K binary classifiers. Some classifiers, such as convolutional networks (defined in Section 5.4, output a vector $h(x) \in \mathbb{R}^K$ where $h(x)_i$ is the probability that x belongs to class i; for binary classification $h(x)$ is reduced to a single real number representing the probability that x belongs to the first class and $1 - h(x)$ represents the probability that x belongs to the second class. Every classifier partitions the *feature space*, the n-dimensional space where features exist, into separate *decision regions*, regions of the space which are assigned the same predicted outcome; a *decision boundary* is the hypersurface that partitions the feature space. Classifiers are sometimes classified as *linear* or *nonlinear* according to the nature of the decision boundary they impose on the feature space. Logistic regression, for instance, is a linear classifier while an artificial neural network (with at least one hidden layer) is nonlinear.

Model selection, selecting the best representation $h(x)$ for a particular problem, equivalently, selecting the best classifier for the problem, is often done via cross validation. *Cross validation* is an statitistical model evaluation technique where each model is trained on a subset of the data set and later validated on a disjoint subset. In *hold-out cross validation* the data set is divided into a training set (usually 60-80%) and a cross validation set, each model is trained using the training set and evaluated on the cross validation set and the model which shows the best performance is selected. *k-fold crossvalidation*, on the other hand, divides the data set in k disjoint subsets (usually 5 or 10) and uses $k - 1$ subsets to train the model and the remaining subset for evaluation, this process is repeated k times for each model leaving out a different subset each time and the k performance measures are averaged to obtain a

final measure for the model. Cross validation is also used to select the *model hyperparameters*, parameters that modify the underlying model, for instance, to select the number of hiddden layers for a convolutional network.

The *cost function* $J(\theta)$ of a classifier measures the amount of error the classifier incurs in for a particular choice of parameters θ. A least-squares cost function for a binary classifier which outputs probabilities (such as convolutional networks) is presented in Equation 1

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 \tag{1}$$

where m is the number of training examples, $h_\theta(x)$ is the output of our classifier with parameters θ and y is the real class of the example. A classifier is trained by choosing the parameters θ that minimize this function, hence, minimizing the expected error of the classifier on the training set. *Gradient descent* is a method used to estimate the parameters that minimize $J(\theta)$; at the start, it initializes parameters at random and iteratively updates each parameter using the gradient of the cost function until it converges to a minimum. Gradient descent is guaranteed to converge to a global minimum if the cost function is convex, convexity of the cost function depends on the model $h(x)$.

The model representation $h(x)$ needs to be chosen carefully. If we have an overly *flexible* model, i.e, when $h(x)$ is a complex function with many parameters to be learned compared to the size of the training set, the classifier will probably *overfit* the data, this means that the parameters are fitted way too closely to the data and will pick up every small fluctuation and noise in the training set. This causes the trained classifier to produce almost perfect results on the training set but perform poorly on previously unseen examples. The opposite is also true, when $h(x)$ is very simple the classifier lacks the power to model the function of interest and we say that it *underfits* the data. This problem is sometimes referred as the *bias-variance tradeoff*. A high variance classifier is prone to overfitting, while a high bias classifier is prone to underfitting.

A popular way to avoid overfitting (and underfitting) is to use a flexible model trained with regularization. *Regularization* modifies the cost function to include a penalty to the complexity of the model, thus selecting parameters that minimize both the training error of the classifier and the complexity of the model. Equation 2 shows the cost function for l_2-*norm regularization*:

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \frac{\lambda}{2m}||\theta||_2 \tag{2}$$

where $||\cdot||_2$ is the euclidean norm of a vector. In addition to reducing training error, minimizing the regularized cost function will shrinken the parameters θ hopefully setting some of them to zero, thus simplifying $h(x)$. The *regularization parameter* λ regulates the tradeoff between less training error and less regularization error. l_1-*norm regularization* or *lasso* is similar to l_2-norm regularization except that it shrinks the l_1-norm of θ instead of the l_2-norm.

To evaluate the performance of a classifier we use a separate set of examples (a test set) which should have not been used for training or cross validation. Classification accuracy is the standard performance measure in machine learning. *Accuracy* measures the proportion of test set examples which are correctly classified. Its compliment, *error rate*, measures the proportion of test set examples which are incorrectly classified. Accuracy, nonetheless, is not a good evaluation metric for unbalanced data sets, data sets which have many more examples

of one class than the other e.g., cancer data sets are often unbalanced as most examples belong to the negative class (no cancer) than the positive class (cancer). For instance, a classifier which always predicts no cancer regardless of the input will show a high accuracy (equivalently a low error rate) even though it is not a good model for the problem.

A different set of metrics based on its confusion matrix are used to evaluate the quality of a classifier in unbalanced data sets. A *confusion matrix* is a matrix which summarizes the results of a classifier in the test set (see Table 1). *True positives* is the number of positive

		Actual class	
		Positive	Negative
Predicted class	Positive	True Positives (TP)	False Positives (FP)
	Negative	False Negatives (FN)	True Negatives (TN)

Table 1: Confusion matrix for a binary classifier

examples which were correctly predicted as positive. *False positives* is the number of negative examples which were incorrectly predicted as positive. True negatives and false negatives are defined in a similar fashion. Based on the confusion matrix we can compute some commonly used metrics:

$$Sensitivity \text{ or } Recall = \frac{TP}{TP + FN} \tag{3}$$

$$Specificity = \frac{TN}{FP + TN} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

Sensitivity and specificity are usually preferred to present results in medical diagnosis meanwhile precision and recall are preferred in machine learning. *Sensitivity* measures the proportion of positive examples predicted as positive and *specificity* measures the proportion of negative examples predicted as negative. *Precision* is a measure of the proportion of examples predicted as positive which are actually positive. A good classifier will have both high sensitivity and high specificity or similarly, high precision and high recall. It is always useful to have a single metric to evaluate classifiers, for example to choose between two models. We show two commonly used in Equation 6 and 7.

$$F_1 \ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

$$G\text{-}mean = \sqrt{Sensitivity \times Specificity} \tag{7}$$

In this thesis, we will generally present results for all this metrics (precision, recall or sensitivity, specificity, F_1 score and G-mean). The metric used when selecting a model influences its characteristics and behaviour, hence, one should put some consideration into choosing it. We favor F_1 score over G-mean because it concentrates on prediction in the positive class (cancer) which is harder to predict and the class we are more interested in. Furthermore, it represents a more balanced tradeoff (an small change in precision is corresponded with a small change in recall) than G-mean where an small change in specificity can be corresponded with a big change in sensitivity.

This section is meant to be a compendium of basic concepts in machine learning, practical machine learning involves many subtleties and implementation details not mentioned here. Notation and content in this section is mostly based on materials from Stanford's Machine Learning course[Ng, 2014].

5.3 Artificial Neural Networks

Artificial neural networks or simply *neural networks* are one of the most popular nonlinear classifiers used today. They were initially inspired by the way biological neurons process information coming from its dendrites and relaying it through its axon to neighboring neurons [McCulloch and Pitts, 1943, Widrow and Hoff, 1960, Rosenblatt, 1962] but evolved to become practical for nonlinear modelling albeit less biologically accurate [Rumelhart et al., 1986]. We discuss here multilayer feedforward neural networks, the name should become obvious after a few paragraphs.

Multilayer feedforward neural networks are composed of L layers of *neurons*, units of computation, each of which is fully connected to the next and previous layer (except for the first and last layer). The first layer, called the *input layer*, has $s^{(1)} = n$ units and receives the feature vector $x \in \mathbb{R}^n$ meanwhile the last layer or *output layer* has $s^{(L)} = K$ units corresponding to the K possible classes (or 1 unit for binary classification). Every other layer is called a *hidden layer* (see Fig. 4 for an example). The neural network receives an input $x \in \mathbb{R}^n$, processes it layer by layer and outputs a vector $h_\Theta(x) \in \mathbb{R}^K$, where $h_\Theta(x)_i$ is the predicted probability that x belongs to class i. Each unit performs a computation on the input from the units in the previous layer and transmits the result to the units in the next layer through its connections. Furthermore, each connection has a *weight* w which is to be learned in the training phase, i.e, the weights are the parameters Θ of the model. A neural network is said to be *shallow* or *deep* according to its number of layers or *depth*.[2]

A unit i in layer l computes a function of the form:

$$a_i^{(l)} = g\left(\sum_{j=0}^{s^{(l-1)}} \Theta_{ij}^{(l-1)} a_j^{(l-1)}\right) \tag{8}$$

where $a_i^{(l)}$ is called the *activation* or output of unit i in layer l; $g(\cdot)$ is an *activation function* (defined below); $s^{(l)}$ is the number of units in layer l, $a_0^{(u)} = 1$, for all $u = 1, \ldots, L-1$ (bias units); $a_v^{(1)} = x_v$ for all $v = 1, \ldots, n$ i.e, the activation of the input layer is the input x, and $\Theta^{(l)} \in \mathbb{R}^{s^{l+1} \times s^l}$ is the matrix of weights connecting layer l to $l+1$. At each layer (except the output layer) we include a unit which always emits activation 1 ($a_0^{(1)} = 1$, $a_0^{(2)} = 1$, etc), these are called *bias units* [3]. The bias units are assumed to be included into each vector $a^{(l)}$, hence the sumation in Equation 8 starts at 0 and not 1. It may seem like a convoluted definition but it simply defines the activation of a given unit as the weighted linear combination of activations of the units in the previous layer passed through a nonlinear function $g(\cdot)$. Lastly, notice that $a^{(L)} \in \mathbb{R}^{s^L}$, the vector of activations in the last layer of the network, is equal to the predicted probabilities $h_\Theta(x) \in \mathbb{R}^K$.

[2] There is no consensus on when a neural network becomes a deep neural network[Schmidhuber, 2015]. We consider networks with over 5 layers to be deep.

[3] They are included for a technical detail: so that the activation function $g(w^T x + w_0)$ can shift in the x-axis changing its threshold to $-w_0$, which is learned by the neural network.

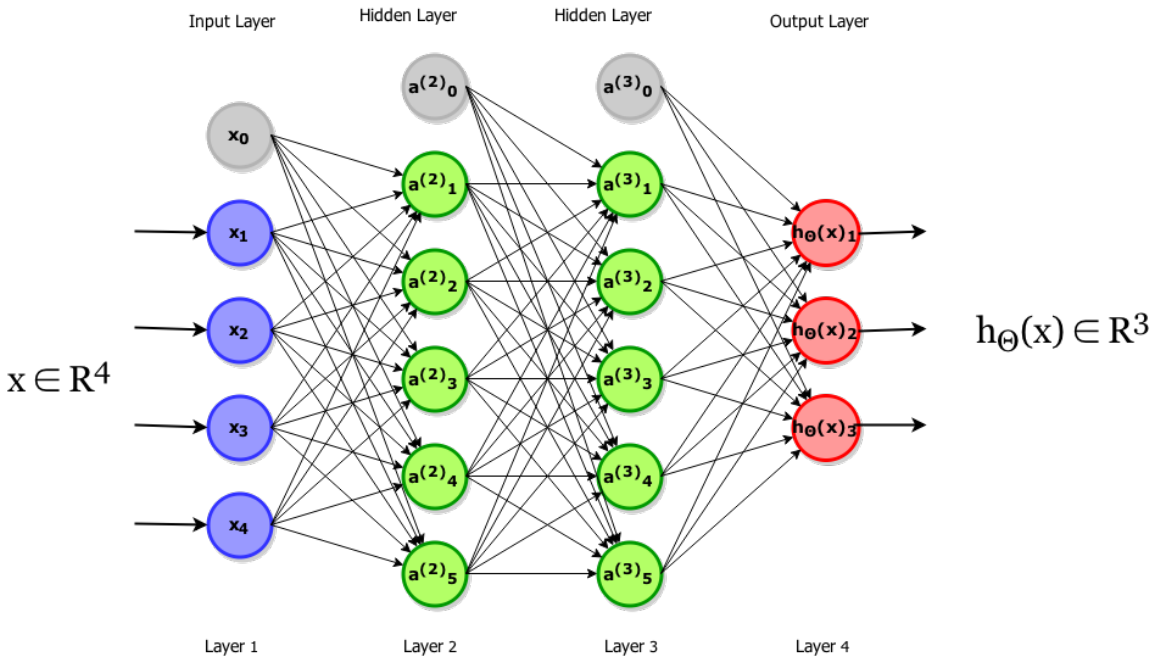

Figure 4: Small neural network example. Input layer with 4 units (blue), two hidden layers of 5 units (green) and output layer of 3 units(red). Bias units appear in gray. It approximates a function $h_\Theta(x) : \mathbb{R}^4 \to \mathbb{R}^3$, i.e., it classifies an input vector $x \in \mathbb{R}^4$ into 3 possible classes.

The activation function $g(\cdot)$ is usually a *logistic sigmoid function*:

$$g(z) = \frac{1}{1 + e^{-z}} \tag{9}$$

The sigmoid function has range [0,1] and is differentiable with respect to z. Because of this characteristics it is used to represent probabilities in the logistic regression classifier. *Logistic regression* for binary classification models the probability that $x \in \mathbb{R}^n$ belongs to the positive class as $g(w^T x)$ and estimates the parameters $w \in \mathbb{R}^n$ during training. Any input whose output $g(w^T x)$ is greater than 0.5 is classified as positive, otherwise it is classified as negative. The sigmoid function equals 0.5 when $w^T x = 0$, thus, the decision boundary of a logistic regression classifier is $w^T x = 0$, which is a linear function.

However, the sigmoid function, per se, is not linear on its input z. Therefore, each unit in a neural network with sigmoid activation functions outputs a nonlinear activation $g(z)$ which in turn is received by units in the next layer, linearly recombined with the activation of other units and passed again through a sigmoid function; these operations are repeated until the input reaches the output layer. As a result, the function calculated by units in the output layer $h_\Theta(x)$ will be highly nonlinear on the original input x. This is the reason why neural networks can model functions which are highly nonlinear and why increasing the number of layers in a neural network increases the predictive power of the model. By the same token, it may be insightful to think of each unit in a neural network as a feature detector (via logistic regression): units in the first hidden layer are trained to activate when simple features are found on the input, units on the second hidden layer activate when a combination of these simple features is present on the input and so on. Thus, the network will learn to detect

the most relevant features for the classification task and as the number of units increases, it learns ever more complex features (granted that there is enough training data).

The cost function of a neural network classifier is defined as:

$$J(\Theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)} log(h_\Theta(x^{(i)}))_k + (1-y_k^{(i)}) log(1-h_\Theta(x^{(i)}))_k)\right] \tag{10}$$

where m is the number of examples in the training set and $(x^{(i)}, y^{(i)})$ is the i^{th} example. $J(\Theta)$ is differentiable with respect to Θ but non-convex, nonetheless, gradient descent usually converges to a good estimate of the network weights [Ng, 2014]. *Error backpropagation* [Linnainmaa, 1970, Werbos, 1974], an algorithm where error terms are computed on the output layer and backpropagated layer by layer as the weights are adjusted, is commonly used for gradient descent training. Given the big number of parameters which need to be estimated, neural networks are susceptible to overfitting. The simplest approach to overcome this problenm is to use regularization. Regularization for neural networks is done by performing gradient descent on the regularized cost function presented in Equation 11

$$\begin{aligned} J(\Theta) = &-\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)} log(h_\Theta(x^{(i)}))_k + (1-y_k^{(i)}) log(1-h_\Theta(x^{(i)}))_k)\right] \\ &+ \frac{\lambda}{2m}\sum_{l=1}^{L-1}\sum_{i=1}^{s^{(l)}}\sum_{j=1}^{s^{(l+1)}}\left(\Theta_{ij}^{(l)}\right)^2 \end{aligned} \tag{11}$$

5.4 Convolutional Networks

Yet to write

5.5 Convolutional Networks applied to Breast Cancer

5.5.1 Related Work

In this section we offer a summary of some of the first work on using convolutional networks for breast cancer diagnosis as well as some of the articles that have influenced this thesis.

[Lo et al., 1995] were the first group to use convolutional networks for breast cancer detection. They used a CNN with two hidden layers to detect microcalcifications. A high sensitivity image processing technique was used to obtain a set of 2104 patches (16 by 16 pixels) of all potential disease areas from 68 digital mammograms; of these, 265 were true microcalcifications and 1821 were "false subtle microcalcifications". Prior to training the CNN, a wavelet high-pass filtering technique was used to remove the background of these images. Each image was flipped over (left-right) and 4 rotations for each the original and flipped images were used for training (0°, 90°, 180° and 270°). The CNN was composed of one input unit (16×16), 12 units in the first hidden layer (12×12), 12 units in the second hidden layer(8×8) and two output nodes(one for YES and one for NOT). The input size (16), number of hidden layers (2) and kernel size (5×5) was obtained via cross validation, altough not many other options were explored: they tried input sizes of 8, 16 or 32, one or two hidden layers and kernel sizes of 2, 3, 5 or 13. The CNN reached 0.87 average AUC when identifying individual microcalcifications and 0.97 AUC for clustered microcalcifications. Only a minimum of three calcifications was considered a detection. Sensitivity and specificity test results were not

reported. This article proved that simple convolutional networks can be efficiently used for medical image pattern recognition.

6 Methodology

In order to achieve the proposed objectives and test our hypotheses we will need to carry out various tasks. We list them here in the order in which we plan to execute them:

1. **Literature review**

   A thorough review of the published work using the databases and resources available in the institution. By the end of this task, a complete theoretical background should be obtained and written. This will also help refine the scope of the project and the experiments to be conducted.

2. **Software review**

   Once a clear idea of what are the possible experiments to be executed, we will need to find appropiate software to perform them. Software for database managing, preprocessing and implementation of different neural networks should be either located or developed.

3. **Database preprocessing**

   We will ready the database images for the experiments; these implies joining different databases, obtaining the required features, preprocessing the images, assigning labels, etc.

4. **Assesing image preprocessing**

   We will train a standard convolutional network with fixed parameters on mammograms with three different preprocessings: no preprocessing, image enhancement using median or gaussian filters and wavelet filtered images. Furthermore, we will train a deeper convolutional network on nonpreprocessed images. We want to answer three research questions: which is the best preprocessing for convolutional networks, is using the best filter significantly better than using nonpreprocessed images and can a convolutional network automatically preprocess the images?

5. **Exploratory experiments**

   We will train standard convolutional networks in two different inputs: small image patches obtained from mammograms and whole mammogram images. We will also train a linear classifier, probably rectified linear units, on the features obtained from a convolutional network trained on the ImageNet database, i.e., we will use an already trained convolutional network instead of one trained specifically in mammograms. Here we will use the image preprocessing technique that showed better results in the previous step. We want to answer two research questions: Can a convolutional network trained on whole mammograms perform as well as one trained on small patches and can we use an already trained convolutional network to classify mammograms?

6. **Model selection**

Using the insights from previous sections and the current literature on convolutional networks, we will select a network architecture along with novel features, preprocessing, training and regularization procedures. We aspire to find the best convolutional network configuration for mammogram classification.

7. **Further experiments**

   We will train the chosen convolutional network on our mammographic database. We will perform crossvalidation to adjust the most important network parameters and use regularization to avoid possible overfitting. We want to answer two research questions: is the performance of the convolutional network considerably improved by parameter tuning and, more importantly, is this a good performance?.

8. **Gathering results**

   Produce results on the test set and elaborate figures and tables. This could be obtained directly from software output or from further program executions.

9. **Reporting results**

   Write the thesis and any article or technical guide which may result from this work. Both this and the previous step will be performed along the execution of the project, hopefully benefiting from the supervisors' feedback.

Finally, we would like to note that this is an idealized workflow and some changes may occur due to time limitations or lack of resources. In the unlikely case that the work is finished before the project deadline, we will either reiterate on model selection, experiments and results gathering and reporting or look into digital tomosynthesis, network ensembles or evolving convolutional networks.

7 Work Plan

We present here the expected work plan for this master's thesis. A description of the activities can be found in Section 6

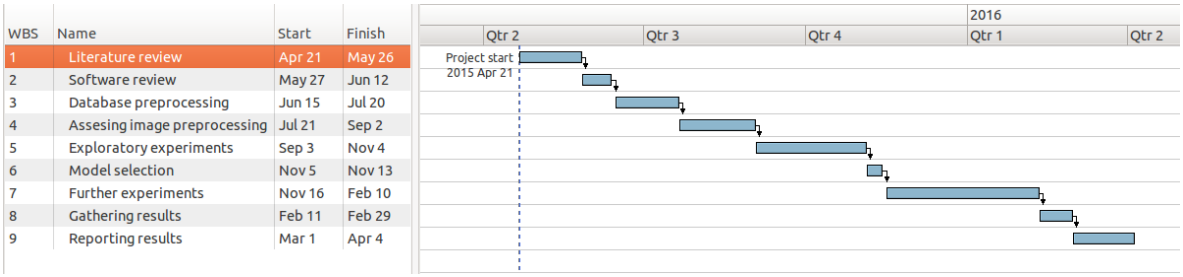

Figure 5: Thesis work plan.

Literature and software review and the preprocessing of the image database are expected to be done before the end of the summer term (late July). The first experiments about preprocessing and using convolutional networks on entire mammograms are expected for the fall semester (late October to early November). At this point, these results should be reported in the thesis. Later, the final architecture and methods will be selected and the final

experiments run. The thesis document should be delivered by the start of March. During this month, if the results are valuable, we expect to write a conference or journal article to share our results with the community.

References

[American Cancer Society, 2014] American Cancer Society (2014). Mammograms and other breast imaging tests. electronic, Atlanta, GA. Available online on `cancer.org/acs/groups/cid/documents/webcontent/003178-pdf.pdf`.

[American Cancer Society, 2015] American Cancer Society (2015). *Cancer Facts & Figures 2015*. American Cancer Society, Atlanta, GA. Available on `cancer.org/acs/groups/content/@editorial/documents/document/acspc-044552.pdf`.

[Breast Cancer Surveillance Consortium, 2013] Breast Cancer Surveillance Consortium (2013). Performance measures for 1,838,372 screening mammography examinations from 2004 to 2008 by age–based on BCSC data through 2009. electronic. Available on `breastscreening.cancer.gov/statistics/performance/screening/2009/perf_age.html`.

[Dieleman et al., 2015] Dieleman, S., Willett, K. W., and Dambre, J. (2015). Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly Notices of the Royal Astronomical Society*.

[Fukushima, 1980] Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36:193–202. PMID: 7370364.

[Hernandez, 2014] Hernandez, J. L. (2014). Selección de características y clasificacíon de masas por medio de redes neuronales, máquinas de vector de soporte, análisis discriminante y regresíon logística en mamografías digitales. Master's thesis, Tecnológico de Monterrey, Monterrey, Mexico.

[Howlader et al., 2014] Howlader, N., Noone, A. M., Krapcho, M. F., Garshell, J., Miller, D. A., Altekruse, S. F., Kosary, C. L., Yu, M., Ruhl, J., Tatalovich, Z., Mariotto, A. B., Lewis, D. R., Chen, H. S., Feuer, E. J., and Cronin, K. A. (2014). SEER cancer statistics review, 1975-2011. review, National Cancer Institute, Bethesda, MD. Available on `seer.cancer.gov/csr/1975_2011/`.

[Kerlikowske et al., 2011] Kerlikowske, K., Hubbard, R. A., Miglioretti, D. L., Geller, B. M., Yankaskas, B. C., Lehman, C. D., Taplin, S. H., and Sickles, E. A. (2011). Comparative effectiveness of digital versus film-screen mammography in community practice in the united states: A cohort study. *Annals of Internal Medicine*, 155(8):493–502.

[LeCun et al., 1998] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

[Linnainmaa, 1970] Linnainmaa, S. (1970). The representation of the cumulative rounding error of an algorithm as a taylor expansion of the local rounding errors. Master's thesis, University of Helsinki, Helsinki, Finland.

[Lo et al., 1995] Lo, S.-C. B., Chan, H.-P., Lin, J.-S., Li, H., Freedman, M. T., and Mun, S. K. (1995). Artificial convolution neural network for medical image pattern recognition. *Neural Networks*, 8(7–8):1201 – 1214. Automatic Target Recognition.

[McCulloch and Pitts, 1943] McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133.

[National Cancer Institute, 2012] National Cancer Institute (2012). What you need to know about breast cancer. electronic, Bethesda, MD. Available online on `cancer.gov/publications/patient-education/WYNTK_breast.pdf`.

[National Cancer Institute, 2014] National Cancer Institute (2014). Mammograms. electronic. Available on `cancer.gov/cancertopics/types/breast/mammograms-fact-sheet`.

[Ng, 2014] Ng, A. (2014). Machine learning course. online. Available on `coursera.org/course/ml`.

[Pisano et al., 2008] Pisano, E. D., Hendrick, R. E., Yaffe, M. J., Baum, J. K., Acharyya, S., Cormack, J. B., Hanna, L. A., Conant, E. F., Fajardo, L. L., Bassett, L. W., D'Orsi, C. J., Jong, R. A., Rebner, M., Tosteson, A. N. A., and Gatsonis, C. A. (2008). Diagnostic accuracy of digital versus film mammography: Exploratory analysis of selected population subgroups in DMIST. *Radiology*, 246(2):376–383. PMID: 18227537.

[Rosenblatt, 1962] Rosenblatt, F. (1962). *Principles of neurodynamics: perceptrons and the theory of brain mechanisms*. Report (Cornell Aeronautical Laboratory). Spartan Books.

[Rumelhart et al., 1986] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning internal representations by error propagation. In Rumelhart, D. E., McClelland, J. L., and PDP Research Group, C., editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1*, pages 318–362. MIT Press, Cambridge, MA.

[Russakovsky et al., 2014] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2014). ImageNet Large Scale Visual Recognition Challenge. *CoRR*, abs/1409.0575. Available online on `arxiv.org/abs/1409.0575`.

[Schmidhuber, 2015] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61(0):85–117.

[Skaane et al., 2007] Skaane, P., Hofvind, S., and Skjennald, A. (2007). Randomized trial of screen-film versus full-field digital mammography with soft-copy reading in population-based screening program: Follow-up and final results of oslo II study. *Radiology*, 244(3):708–717. PMID: 17709826.

[Taigman et al., 2014] Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). DeepFace: Closing the gap to human-level performance in face verification. In *Computer Vision and Pattern Recognition (CVPR)*, Columbus, Ohio.

[Werbos, 1974] Werbos, P. J. (1974). *Beyond regression: new tools for prediction and analysis in the behavioral sciences*. PhD thesis, Harvard University, Boston, MA.

[Widrow and Hoff, 1960] Widrow, B. and Hoff, M. E. (1960). Adaptive switching circuits. In *1960 IRE WESCON Convention Record, Part 4*, pages 96–104, New York. IRE.