Université Pierre et Marie Curie - Paris VI
Télécom ParisTech

Master 2 of Informatics - Imaging

Internship Report

DETECTION OF CANCEROUS TISSUE IN FULL FIELD OCT IMAGES USING CONVOLUTIONAL NEURAL NETWORKS

Author:
Diana Mandache

Supervisor:
Vannary Meas-Yedid

September 2017

Acknowledgments

First of all I would like to express my gratitude and most sincere appreciation to my kind supervisor, Vannary Meas-Yedid, who attentively guided my research while also allowing me to follow my intuition. Furthermore, she felt when I needed encouragement and her moral support gave me confidence and motivated me a lot.

Special thanks to Jean-Christophe Olivo-Marin, head of the Bioimage Analysis Unit at Institut Pasteur, for welcoming me into his team and inspiring both humanity and professionalism.

I need to acknowledge the involvement of LLTech, through Eugénie Dalimier who offered insight into a domain less known to me and received my work with an open spirit. Moreover, I thank Bertrand Le Conte de Poly, the CEO of LLTech, for trusting my scientific skills and accepting to continue the collaboration in form of a CIFRE doctorate.

I would like to thank my colleagues for warmly adopting me into the lab and always creating a positive atmosphere.

Last but not least, a big thank you to my parents whose moral support overcame the physical distance.

Abstract

This paper presents the work I did during a five month internship at the Bioimage Analysis Unit of Institut Pasteur, that marks the end of my master studies in Informatics and Imaging at Université Pierre et Marie Curie (Paris). The purpose of this project is to exploit techniques of Deep Learning in the biomedical context, more specifically, to automatically detect cancerous areas from skin excisions imaged with Full Field Optical Coherence Tomography (FF-OCT) scanner. The objective is to extract representative features from these images and, on their basis, to build a classifier that learns a generalized distribution of the data.

Résumé

Ce rapport présente le travail que j'ai effectué pendant mon stage de cinq mois au sein de l'Unité d'Analyse d'Images Biologiques, à l'Institut Pasteur, marquant la fin de mes études en Master d'Informatique et Imagerie à l'Université Pierre et Marie Curie (Paris). L'objectif de ce projet est d'exploiter les techniques de Deep Learning dans le contexte biomédical, et plus particulièrement pour détecter automatiquement les zones cancéreuses sur des images d'échantillons de peau obtenues avec un scanner de Tomographie par Cohérence Optique plein champ (FF-OCT). Le but est d'extraire des caractéristiques représentatives pour ces images et, basé sur elles, de construire un classificateur qui apprend une distribution généralisée des données.

Contents

1 Introduction

1.1 Motivation

According to the American Cancer Society [12] skin cancer is the most common malignancy, mostly represented by non-melanoma types: Basal Cell Carcinoma (BCC) and Squamous Cell Carcinoma (SCC), from which 80% are BCC. There are about 5.4 million cases of such cancers each year in the U.S. from 3.3 million patients (a person usually develops multiple cancers). The number of cases has been growing, however death from this type of skin cancer is unusual, there are about 2.000 deaths yearly in the U.S. caused by a very weak immune system.

The ultimate goal of this project is to speed up Mohs Surgery which is the preferred procedure (gold standard) for treating the non-melanoma skin cancers.. This technique's high rate of success (>98%) relies on its carefully controlled approach which consist in progressive (gradual) removal of thin layers of tissue. As each layer is extracted, its margins are examined for the presence of cancer, this guides the extraction of the next layer. The examinations are the most time consuming step (on average 2-3 hours per surgery) because the sample needs to be frozen, sliced, stained and only then microscopically analyzed. In the meantime the patient is temporarily bandaged and he has to wait [11].

Patient's comfort can be increased and surgery time significantly reduced by replacing classical histology with the so-called optical biopsy. This refers to non-invasive imaging methods that use the properties of light to visualize the tissue at different depths. Instead of being mechanically cut, the sample is imaged at different levels on the Z axis, allowing for a 3D view assembled from a stack of 2D slices.

Histology slide scanner microscopes have a resolution of less than a micron, so the main requirement of an optical biopsy system is to provide cellular-level resolution while assuring enough penetration depth. The technologies generally employed in this domain are optical coherence tomography (OCT) and fluorescence confocal microscopy, however we will use Full Field Optical Coherence Tomography (FF-OCT) which is demonstrated to surpass them, offering the best balance between the two before-mentioned properties: resolution and penetration depth. The theory of FFOCT will be detailed in section 2.

Since FFOCT is a state of the art technique there's a high necessity in training pathologists, so in the purpose of making it more accessible, we intend to develop an automatic aid-to-diagnosis tool. To achieve this we rely on the advancing theory of Deep Learning which recently became the top solution for tasks like object detection, face and speech recognition etc, the kind of tasks that "come naturally" to humans, but are difficult to define algorithmically.

There is a growing tendency for medical diagnosis applications to rely on deep learning which is also fueled by the fact that this technique tends to win the grand challenges in the domain: Tumor Proliferation Assessment Challenge - MICCAI '16, Breast cancer metastases detection in lymph nodes - ISBI '16 '17, all having as data histology slides. Recently, neural networks conquered the field of dermatology, with the method of Esteva [3],validated by 21 pathologists, that classifies cancerous lesions from macro images of the skin surface. In OCT imaging, deep learning is used for segmentation of the retinal layers [7], however there is almost no research on analyzing Full Field OCT optical biopsies using neural networks.

1.2 Entities involved

The laboratory that welcomed me during the internship is the Bioimage Analysis Unit of the Pasteur Institute where I could benefit from their vast expertise in developing methods of computer vision and image analysis for the biomedical domain. Through the lab's collaboration with LLTech SAS, who produces the Light-CT scanner - an imaging system used for optical biopsy, I was introduced to the FFOCT imaging technique. Their scanners are spread all over the world, one of them being at Drexel University Medical College, in Philadelphia, USA which provided us with both images of surgical excisions and their interpretation *i.e.* diagnostic. The communication between us and the pathologist was possible through Cytomine [16] which is an open-source web application developed by the University of Liège, it is intended for big image management and annotation inside multidisciplinary teams for collaborative projects.

Bioimage Analysis Unit, Pasteur Institute

The Pasteur Institute of Paris needs no introduction, since its creation in 1887 it is fighting against infectious diseases together with its international network of 29 institutions spread on 5 continents. It managed to become a world renowned institution not only through the name of its founder Louis Pasteur - the first Professor of Microbiology and the founder of Immunology, who made some of the greatest breakthroughs of his time, like pasteurization and vaccines for anthrax and rabies, but also for the 8 Nobel prizes awarded since the 1900's, including the one in 2008 for the early discovery of HIV.

Bioimage Analysis Unit is one of the 18 teams of the Cell Biology & Infection Department which brings together biologists with computer scientist, physicists and mathematicians in order to develop software tools for biological image analysis and quantification. Among the recent interests of the lab are:

multi-particle (viruses, genes, etc) detection and tracking by Bayesian filtering and Wavelet transform, cell segmentation by active contours and analysis of histological images for digital pathology. All these applications are made available to the biological community under one software that reunites them, the *Icy* platform [15].

Besides the knowledge available in the lab, I could profit from the huge computational resources of the institute, which it is known to be crucial in efficiently developing Deep Learning algorithms. Besides hundreds of CPUs and thousands of gigabytes of RAM, the Pasteur computational cluster has GPU nodes formed of last generation Nvidia Tesla GPUs which are especially dedicated to training deep learning models. For example the M40 model is claimed to be "the world's fastest deep learning training accelerator" with a speed of 7 teraflops (7 trillion floating point operations per second), memory of 24 GB and bandwidth of 288 GB/s.

2 Full Field Optical Coherence Tomography

2.1 Context

There is a plethora of non-invasive imaging techniques, whose field of application is determined by the trade-off between resolution and penetration depth: going from magnetic resonance imaging (MRI) or computed tomography (CT) used in studying organs anatomy (big depth full body, big resolution of the order of millimeters) to optical coherence tomography (OCT) and confocal microscopy. OCT provides higher resolution (few microns) while offering a shallow penetration depth (≈1-2mm), making it best suited for in-vivo study of the layers composing the retina (back of the eye), or endoscopic imaging of arteries, esophagus, intestines[6]. On the other hand, confocal microscopy offers an excellent lateral resolution (0.8 µm), but an insufficient axial resolution which doesn't give enough penetration for executing an optical slicing comparable to histology.

Full Field Optical Coherence Tomography (FFOCT), developed by the ESPCI team of Pr. Claude Boccara and made commercially available by LLTech in 2011 for research purposes in biological tissue morphology and function, comes to fill a gap between classical OCT and confocal, offering a resolution of 1 µm in all 3 dimensions.

2.2 Principle

FFOCT relies on the same principle as OCT, the property of interference of light, this is a phenomenon in which two waves superpose to form a resultant wave. It usually refers to waves that are correlated (coherent) to each other because they belong to the same source. For the sake of a full explanation of

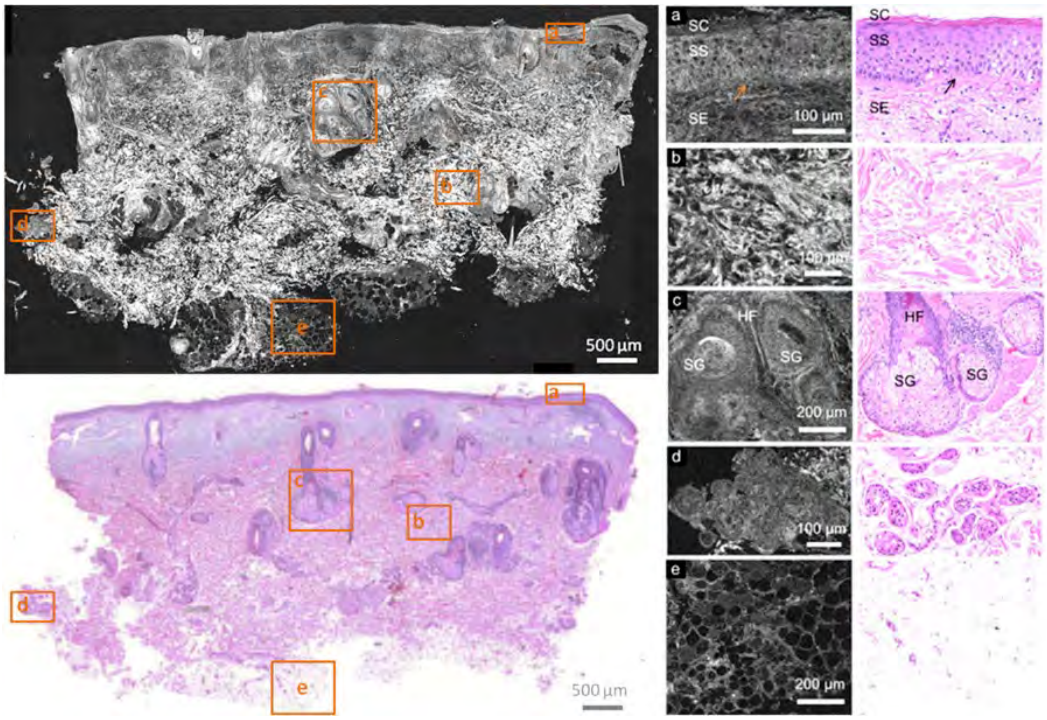


Figure 1: Normal skin morphology: FFOCT vs. Histology: a. Epidermis b. Collagen c. Hair follicle + Sebaceous Glands d. Sweat glands e. Fat cells

the denomination Optical Coherence Tomography, tomography means imaging by sections or sectioning, through the use of any kind of penetrating wave: light, sound, x-ray etc.

In a FFOCT system the light from a white halogen lamp is broken by a beam splitter into two arms: a sample arm containing the examined sample and a reference arm containing a mirror, then it combines again and it is captured by the camera. By moving the reference arm, one can image the sample at different depths. The obtained image is a result of the optical properties of the tissue under investigation, such as differences in refractive indexes or scattering variations and, to a lower extent, differences in absorption.

As opposed to classical OCT, which performs a single-point raster scan producing a cross-section XZ image, FFOCT produces "en face" XY images because it has an array detector i.e. megapixel camera (CCD or CMOS). This explains the name "full field", because it captures the full XY plane at once, instead of point by point.

2.3 Specifications

The Light-CT scanner that is based on the FFOCT technique is able to perform microscopic sub-surface visualization of tissue, without damaging it in any way (no cutting, no markers, no photo bleaching). It offers a cellular-level resolution of 1 micron in 3D and a speed of scanning of 4-5 minutes per cm^2 with a supported sample size of maximum 2.7cm in diameter (corresponding to around $7cm^2$) and 5 mm in height. The penetration depths it can reach depend on the type of tissue examined, it can vary between 200 µm and 1mm It has a simple setup, the scanner has the size of a regular microscope and it just need to be connected to a computer whit its dedicated software installed on it.

2.4 Dynamic FFOCT

A new acquisition method has been developed recently by [5], it is destined to work on the same setup but it also quantifies microscopic movements. It brings information on the metabolic activity at an intracellular scale that is complementary to the morphological information.

By observing the sample for a given duration and subtracting the information of the first "frame" taken as reference, the changes in the tissue can be measured. The meaningful information is extracted from the raw signal by calculating the standard deviation of the signal in time.

FFOCT produces grayscale images, which they are known to be harder for the human eye to discern, also there is speckle noise present which makes it almost impossible to remove in post-processing without losing some information that is hidden in the noise. What is more, certain structures of the tissue can overshadow others because of their physical (optical) properties, for example cancerous cell nuclei can become less visible if they are surrounded by an abundance of collagen whose refractiveness makes it produce strong signal, appearing very bright. Nevertheless, D-FFOCT solves this problem, because it makes only the cells visible because collagen fibers are static. What is more, both spatial and temporal information can be superposed in one image, in color this time.

3 Fundamentals of Deep Learning

3.1 Context

Classical algorithms represent an explicit solution to a well-defined problem that can be translated in computer language under the form of specific instructions. However some tasks are ambiguous to define in such a way, ironically enough, those tasks are the ones that "come naturally" to humans, like: object detection, speech or handwriting recognition.

From the idea of mimicking the human reasoning (decision making mechanism) a new field of computer science emerged in the 1950s: **Artificial Intelligence**. At the beginning intelligence consisted more of "knowledge", under the form of specifically organized databases on which there were applied logical rules. The most popular success of AI was in 1997, when DeepBlue, developed by IBM, beat at chess the world champion, by testing the outcome of all possible moves. There also existed some attempts in the medical domain, like ONCOCIN [10] used by Stanford's Oncology Clinic in the 80s. It was an expert system that based on an extensive history of cancer cases chose the chemotherapy plan with the best chance of cure and the least chance of side effects.

Later, in the 1980s, the concept of learning was introduced in the form of **Machine Learning**: algorithms based on statistics, probabilities and optimization can solve problems by experience, studying an extensive set of examples and finding patterns in them, then use that prior training on new data.

There are two types of learning: supervised and unsupervised. In a **supervised** problem, the input data is accompanied by the desired results and the algorithm has to learn a generalized rule that maps inputs to outputs: if the

outputs are categories/classes then it's a classification problem, and if the outputs are continuous then we call it a regression problem. In **unsupervised** learning, patterns in the data are found with no guidance (no given outputs), the main problem being clustering - splitting data into groups. Some popular methods are: random forests, support vector machines (SVM) (supervised), k-means (unsupervised).

In computer vision, it's not very effective to consider the attributes of the data to be the raw pixels, but to define some **features** which encode the data; this can also be seen as a dimensionality reduction problem. A popular approach is **bag-of-features**[14]: 1) from a set of images extract salient keypoints (SIFT[18], SURF[19], HOG[20]), 2) split keypoints in k groups by k-means clustering, 3) only one keypoint is enough to represent one cluster - the center of the cluster, 3) all the cluster centers (which are nothing else but patches of the images: can be textures, small object parts etc) form a vocabulary, 4) encode an image using the vocabulary, 5) train a classifier (e.g. SVM) in the space of the vocabulary. Another approach is to have already defined so-called **filter banks**, independent from the dataset [21]. A filter convolved with an image gives a feature map which brings lower level information like the presence of edges with a certain orientation. Here learning consists in discriminating the data according to their responses to the filters.

We notice that a crucial part in a good machine learning algorithm is finding the most suitable definition for the attributes of the data according to the problem, this is where **Deep Learning** steps in.

Deep learning has its roots in 1957 with the invention of the **perceptron** [17], which served as a linear classifier, but it is basically the prototype of the modern artificial neuron. However, back then neural networks didn't manage to perform better than existing machine learning algorithms, so research in the domain almost stagnated. In 1986 a major discovery was made, the

backpropagation[24] algorithm which was used for adjusting the weights proportionally with their influence in the error. Thanks to it, in 1997, LeCun [28] managed to develop the first large-scale practical application of (convolutional) neural networks ***LeNet***: handwritten digits recognition on the MNIST database.

The field of deep learning developed slowly (and networks got deeper) together with the improvement of **GPUs** until the major boom of 2011 when Cireșan and colleagues [27] achieved superhuman performance in several image recognition challenges with deep neural networks, "AlexNet"[25] surpassed classical machine learning approaches by a significant percentile, winning the ImageNet competition (1.3 million high-resolution natural images from 1000 classes). Furthermore, in the medical domain Cireșan's neural networks [26] also outperformed other approaches, in the ICPR'12 and MICCAI'13 challenges for breast cancer (mitosis) detection in large histology images.

3.2 Principle

Deep learning is synonym with deep **neural networks**, which represent a biologically inspired computational model that consists of interconnected layers of artificial neurons. The output of one neuron represents the input of one or more neuron from the next layer and there are no connections between neurons of the same layer, so a neural network is an acyclic graph. A neuron can serve as a stand alone weak classifier, so a network can be seen as a combination of weak classifiers that form a strong one. The *intelligence* of a network resides in the weights of its neurons, which also represent its adjustable parameters. To get to correctly solve a task, the network adjusts its parameters through a learning mechanism that is based on *trial and error* principle.

A **neuron** has more inputs and one output value, each input value is scaled up or down according to its corresponding weight and then the sum of the products is fed to an activation function which mimics the biological excitation of a neuron. Theoretically, an on-off switch is best modeled by the Heaviside *a.k.a* step function which is 1 if its input value is positive and 0 if else, however, in nature, the change of state does not happen instantly so the Sigmoid *a.k.a* soft step function was introduced $f(x) = \frac{1}{1+e^{-x}}$. In modern architectures, rectified linear logic unit (ReLU) is generally used because it accelerates the convergence of the learning mechanism to a factor of 6 [25] and because of its simpler definition $f(x) = max(0, x)$, so it is just a threshold that removes weak signals.

Neurons are interconnected, the output of one transfers to the input of another through $\hat{y} = f(\sum_i w_i x_i + b)$ called **forward propagation**. As an input travels through the layers of neurons it produces an output that is compared with the ground truth (the expected output) through a **loss function** which computes the error between the expected and the predicted output. The loss function usually used is cross entropy $L(y, \hat{y}) = -\sum_i y_i log(\hat{y}_i)$. The goal is to minimize this error as much as possible and the way to do it is by adjusting the weights accordingly. This adjustment is done with the **back-propagation** procedure which distributes the prediction error (i.e. loss) to all interconnected neurons in a path, proportionally to their contribution to the error. Mathematically, it means computing the gradient (or derivative) of the loss function with respect to the weights of a multilayer stack of neurons, so applying the *chain rule* for derivatives: $\frac{\partial c}{\partial a} = \frac{\partial c}{\partial b}\frac{\partial b}{\partial a}$. Then, with the derivative of the loss computed with respect to each weight in the network, the weights are adjusted in the negative direction of the gradient (or derivative) with a step called *learning rate*: $w_i \leftarrow w_i - \eta\frac{\partial L}{\partial w_i}$.

3.3 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are designed for matrix inputs, namely images. They are inspired by the visual cortex of some animals, every neuron responding to a small area of the field of view called receptive field. A **convolutional layer** is composed of more **feature maps**, which represent the repose of the input after being convolved with a **filter**, a filter defines a feature and the map signals its presence in the input image. A filter represents the weights of the neurons and the feature map their outputs, the inputs of the neurons are connected to patches of either the input image or the feature maps of the previous layer, these patches correspond to the receptive fields.

After a convolutional layer, there is usually a **pooling layer**, it downsamples its input by either taking the maximum value in an area (max-pooling) or taking the mean value (average-pooling). The motivation behind downsampling is to reduce the number of parameters of the network, reduce the representation size of data and makes the network less sensitive to shifts or transformations.

After several convolutions and downsamples which serve as feature extractors, one ore more fully connected layers are added. As the name suggests, their neurons are connected to every neuron from the previous layer and to the next layer. They can learn non-linear combinations of the features discovered before and they serve as a classifier.

CNNs exploit the property that images are compositional hierarchies, hence lower layers learn basic features (e.g. edges, blobs) and higher layers learn a combination of the features from previous layers (i.e. textures, object parts).

4 Application

4.1 Data

The quantity and also the quality of the data are crucial for a machine learning algorithm to perform satisfactory. For the beginning of this project, the quantity of data was limited: 10 images manually segmented by a pathologist and annotated with the labels BCC and normal, only 3 of them contained cancerous (BCC) tissue. Out of the whole annotated data, 93,8% was normal tissue and only 6,2% cancerous zones. Besides the massive imbalance in the amount of data in each class, the intra-class variation is also important, as it can be seen in Fig.1 a normal skin sample imaged with FFOCT presents multiple distinctive structures like: epidermis, sebaceous and sudoriparous glands, fat cells and collagen, while cancerous regions have very similar appearance: high density of dark nuclei of around 10px diameter. On average, an image has 200 Megapixels and we build the dataset by splitting them in patches of 200x200 pixels; this is a suitable size for both appropriate context capturing and respecting memory constraints imposed by deep learning methods.

Deep learning algorithms don't perform well on unbalanced datasets, this can also be deduced from the fact that most datasets used for academical purpose are balanced, whereas real-life problems are anything but balanced, like in the medical field: the number of healthy patients massively outnumbers the sick cases. From the tests we conducted on the raw data (non-overlapping 200x200 patches) we concluded that indeed, the majority class was favored, so most of the patches were classified as normal.

Solutions for tackling **class imbalance** problem:

- *more data*: getting more data is expensive and time consuming, especially in the medical field;
- *more categories*: split the more populated class into more subclasses, e.g. label every normal skin structure, but defies the purpose of this project;
- *sampling*: at first sampling was done naively i.e. just randomly discarding samples from the normal class to reach the same number of data in both classes. Then, inspired by the oversampling technique found in the literature which consists in either duplicating or synthetically generating data from the minority class [22, 23], we changed the method of selecting the patches. Our method consists of splitting the normal areas in non-overlapping patches and for the cancerous class we select overlapping patches such that we get around the same amount of samples as for the majority class, this implies choosing the sampling stride size accordingly;
- *negative mining*: smartly discarding samples from the majority class by keeping the most difficult samples i.e. the ones that are misclassified; this technique will surely be tried out in the future;
- *weighting*: samples from minority class influence more the loss function which quantifies the error of the system; from our tests this gives good results only for small ratios (e.g. 1 to 1,5), not 1 to 38 which would correspond to our data proportionality (1.139 BCC patches and 43.289 normal patches, meaning that there is one cancerous data sample for 38 healthy samples, intuitively, one BCC patch has the importance or weight of 38 normal patches). In the end we made use of this technique but only after we augmented the minority class through oversampling (28.396 BCC, 43.289 normal $\Rightarrow 1 : 1.5$ ratio).

As new annotations were received from the pathologist and the dataset was complete (40 images of which 10 cancerous), the classes got even more unbalanced (BCC is less than 1% of the data) and new characteristics appeared for the BCC areas. One explanation we discovered was that there was more collagen present and because it is more reflective than the cancerous cells they become obscured by it. In some cases, abnormal tissue which seems to "swirl" around cancerous zones can be hinting for the presence of BCC. Because this kind of abnormal area was not labeled, we tried to capture more context for the BCC class by slightly changing the sampling: before we considered as BCC a patch which had 80% of its area labeled accordingly, but then we opted for just 30%.

For training the network 80% of the data was used, while the rest of 20% was used for testing. Following the last sampling technique mentioned, the first dataset comprised of 10.641 distinct normal patches and 9.741 overlapping BCC patches (out of 98 distinct patches; stride of 20px) and finally, the complete dataset had 43.289 non-overlapping normal patches, 28.396 overlapping BCC patches (out of 1.139 distinct patches; stride of 40px). Also, some data augmentation was performed, by horizontal and vertical flipping, slight rotations and shifts.

4.2 Supervised approach

After testing different state-of-the-art architectures like VGG16 or InceptionV3, pre-trained on ImageNet database or not, we didn't obtain accuracies over 60% so we inferred that those architectures were too deep and complex for our data distribution, however further investigation is intended. The proposed architecture has the simplicity of LeNet combined with ideas introduced by AlexNet and adopted also by VGG, like:

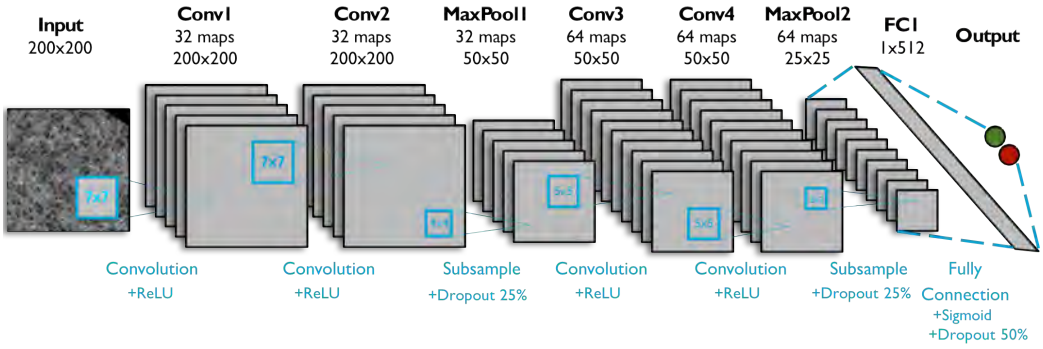


Figure 2: Proposed CNN architecture for supervised learning

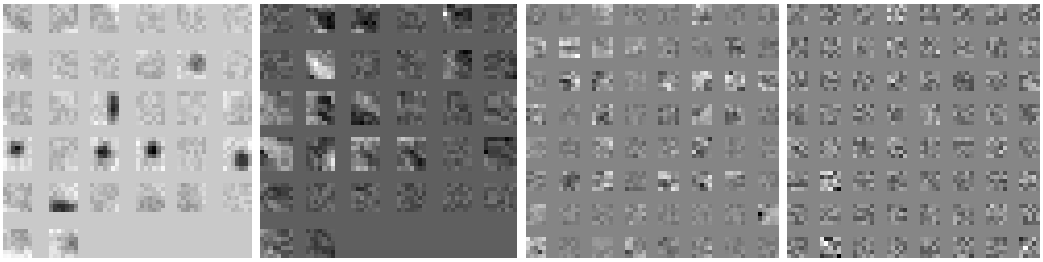


Figure 3: Convolutional filters learned supervised

- convolutional blocks: consecutive convolutional layers (instead of one conv layer followed by pooling) to capture larger input with a save in parameters;
- dropout [29]: after the pooling layers a fraction of neurons is randomly removed; this is done in the learning phase only, with the aim of avoiding overfitting, in other words, generalizing rather than memorizing;
- rectified linear unit (ReLU): activation function used to ease the computations while remaining biologically accurate $f(x) = max(0, x)$.

We trained a 7 layer neural network (see Fig.2): 2 convolutional blocks with 2 convolutional layers each, followed by maxpooling and dropout after every block and finally, a fully connected layer with 512 neurons and 50%

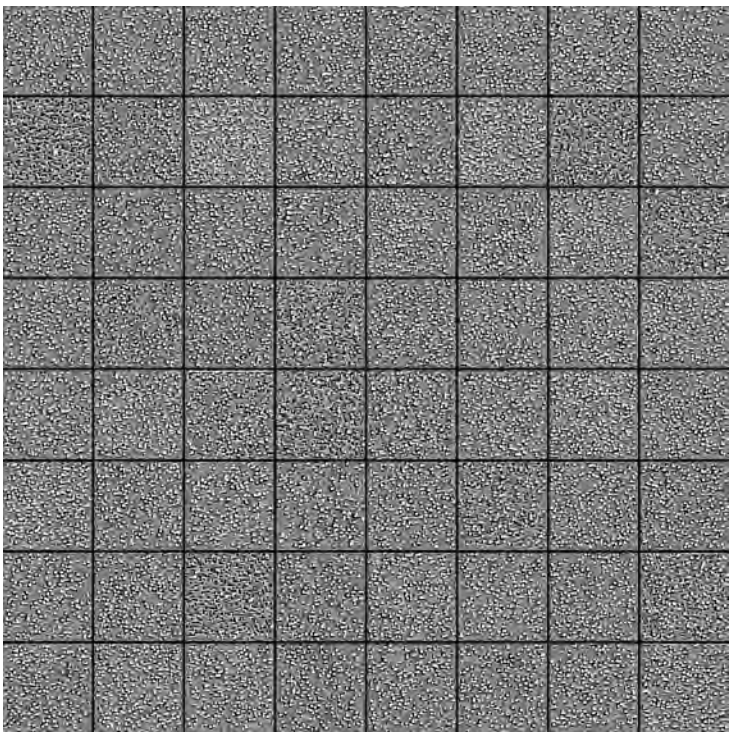
dropout serving as a classifier, followed by 2 output neurons corresponding to each class. The layers from the first block have 32 filters of 7x7px each and the other one has layers with 64 filters of size 5x5px. The network has in total 20.686.561 weights to train. The weights are initialized using the Glorot method [31] which comes from the idea that the gradients of each layer should follow more or less the same distribution at the beginning of training. The aim is to optimize the objective function, which is binary cross entropy, while maximizing sensitivity (*a.k.a.* recall), this was possible using Adam [32] stochastic gradient decent on mini-batches of 32 samples along 200 iterations. Adam (Adaptive Moment Estimation) is one of the adaptive methods of gradient descent (along with Adagrad, Adadelta, RMSprop and more) whose particularity is that they adapt the learning rate to the parameters, performing larger updates for infrequent parameters and smaller updates for frequent ones. The mini-batch gradient descent approach is a trade-off between computational accuracy and convergence time, so between batch (entire dataset) and stochastic (one example at a time) gradient descent. From practice we discovered that a smaller batch-size (i.e. 32 samples) gives better results and from [33] we see that local minima have a better generalization capacity than global minima (offered by computing the gradient on the whole dataset). Training time for the small dataset was around 5 hours on the GPU cluster presented in section 1 for 200 epochs (1,5 mins per epoch); an epoch represents one forward and one backward pass for all the training examples so for all the mini-batches. However, testing is instantaneous on any king of system configuration.

Firstly, training was done on the 10 image dataset, which, as we detailed before, was stable from the point of view of the BCC class. We obtained good results, an accuracy of 94,6% (see column 1 of Table 1) and from analyzing the filters and some activation maps we conclude that, indeed, the network learns to discriminate the BCC class by its concentration of small dark blobs. As we tested the complete 40 image dataset with this network (see column 2 of Table 1) it reinforced this belief. After training on all the data we obtained

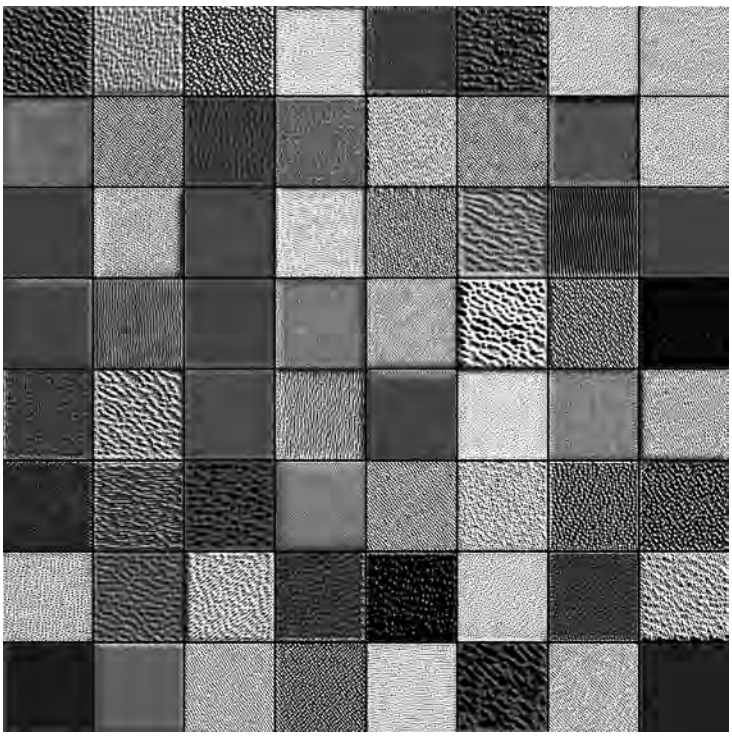
	Train10 & Test10	Train10 & Test40	Train40 & Test40
Specificity	89,26 %	93,22 %	79,48 %
Sensitivity	99,94 %	35,40 %	93,22 %
Accuracy	94,60 %	91,74 %	84,92 %

Table 1: Performance measures for the proposed architecture trained and tested on 10 and 40 image datasets

satisfactory results, an accuracy of 84,92% (see column 3 of Table 1), but the learned filters and the activation maps didn't give much intuition. From comparing the maximum activations (see Fig.4) it is clear that the network trained on 40 images finds more complex textures than the CNN trained on 10 images that focuses on the distribution of blobs of a size of approximately 10 pixels.



(a) trained on 10 images



(b) trained on 40 images

Figure 4: Max input activations of Conv3 layer

4.3 Unsupervised pre-training approach

There is a strong motivations for adopting the unsupervised approach: there is not enough annotated data and one of the goals of the project is to involve pathologists as little as possible so we can pre-train on unlabeled data and fine-tune on our small dataset. With unsupervised learning we can even hope to find hidden patterns in the data that can help us understand the morphology of cancerous tissue in FFOCT imaging, that can even have clinical relevance.

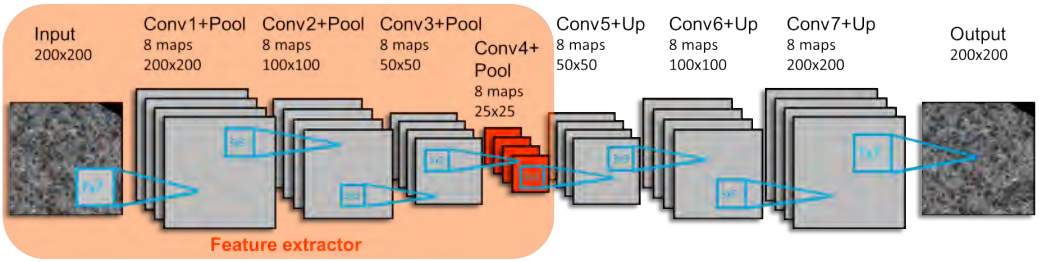


Figure 5: Proposed CAE architecture for unsupervised learning

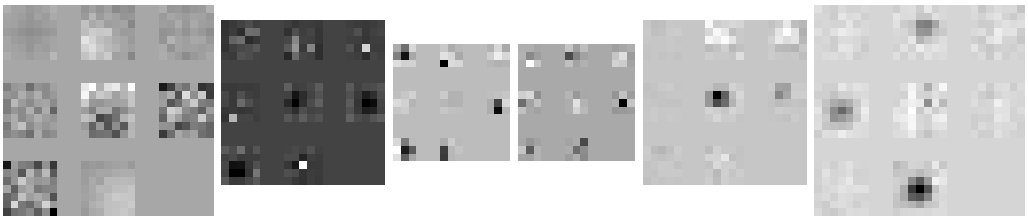


Figure 6: Convolutional filters learned unsupervised

Convolutional Auto-Encoder (CAE) [30], which have architectures similar to CNNs used for segmentation, like Unet [34], but unlike them, it tries to reproduce the input itself, instead of obtaining a segmentation map. CAEs learn to encode and also decode the data with two mirrored networks: one with convolutional and pooling layers that compresses the data and the other with deconvolution (*a.k.a.* reversed convolution) and unpooling that decompresses the data. The data representation can be either compressed or sparse.

The proposed architecture has two symmetrical networks with 3 block of convolution/deconvolution followed by pooling/unpooling (see Fig.5), each convolution/deconvolution layer learns 8 filters of size 7x7, 5x5 or 3x3, then pooling reduces or increases the size for a factor of 2x. The dimensionality of the data is reduced from 40.000px to 5.000px, the input being represented with 8 feature maps of 25x25px at the core of the network. Since the reconstruction is not 100% accurate to the input, the CAE can double as denoising method (see Fig.7).

To use the trained CAE as feature extractor the decoder network is removed from the architecture, leaving only the encoder on top of witch a classifier is added. Several classifiers were tested:

- SVM: training a SVM took 12 hours and gave no results (all examples were classified as BCC);
- Random Forest: training took between 5 and 15 minutes depending on the number of decision trees in the forest, it gave comparable results with 100 or with 500 trees, even if the sensitivity seems good, the overall performance is not satisfactory (see column 1 of Table2);
- Fully Connected Neurons: by adding two layers of fully connected neurons on top of the trained network we built an architecture similar with the one used for supervised learning; the newly added layers have 1024 and 256 neurons, respectively.

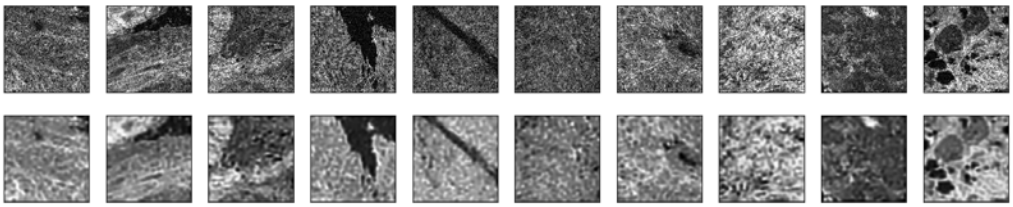


Figure 7: CAE: Original input (top row) *vs.* Decoded output (bottom row)

4.4 Performance measure

In theory, assessing the performance of a classification algorithm is simple and straightforward: simply measuring the percentage of correctly classified data. But if it is obtained an accuracy of 90% when 90% of the data belongs to one class, then the real efficiency of the method is probably 0%. In other words,

	Random Forest 500 trees	Fully Connected 1024 + 256 neurons
Specificity	55,28 %	74,54 %
Sensitivity	90,38 %	97,75 %
Accuracy	73,66 %	86,69 %

Table 2: Performance measures for the proposed pre-trained feature extractor combined with RF and FC classifiers, tested on the 40 image dataset

measuring the performance should be strongly dependent to the problem statement and, especially with neural networks which are black-box models, concluding that a method is correct and robust there is needed some intuition in the reasoning that the method learned.

First of all, taking a look at our specific problem, which is medical diagnosis, what has the greatest importance is the correct classification of the cancerous class rather than the normal class, because it is more grave to send home a sick patient than double check a healthy one. To quantify this we have two measures: sensitivity (true positive rate) that determines the percentage of sick cases identified as such and specificity (true negative rate), measuring correctly classified healthy cases. For this project, having a good sensitivity rate is a priority. This values are obtained from the confusion matrix which shows the count of true positives, true negatives and, more important, false positives (false alarm) and false negatives (miss).

On the other hand, statistical results may be misleading, we can not interpret the efficiency of the method or its behavior when given different data, especially since neural network are a black-box model. In order to get more insight into the reasoning learned by the network we can try to make some sense out of visualizing the network. One thing that can be done is visualizing the weights of the neurons which correspond to the convolutional filters applied on the image to look for certain features; since the filters are usually

very small (e.g. 5x5px) the texture aimed at may not be very obvious to the human eye (see Fig.3 and Fig.6). Therefore, to get the texture that a neuron (filter) is looking for we can visualize the input that would maximize the activation of that neuron, this is done by doing gradient ascent in the input space with respect to the filter activation loss (see Fig.4). Lastly, we can get some intuition by looking at the activation maps *i.e.* the results of a certain input images after being convolved with a filter or, in other words, the transformation of the input as it travels down the network.

5 Conclusion

This project proved to be a crossover from academic theory for solving real world problems in the field of biomedical research. Even if the results obtained are only preliminary, they point towards a promising direction, which is analyzing Full Field OCT images with the powerful methods of deep learning in the purpose of developing computer-aided diagnosis tools and easing its insertion in the clinical environment. Using neural network was motivated by the necessity of discovering features that can best represent this unexplored imagery and we approached this problem both in a supervised and unsupervised way. Hence, we showed that unlabeled images can also be exploited for an improved classification of normal and cancerous tissue. This is relevant because it proves that data can also be "blindly" analyzed and only a minimum expertise of pathologists is required. This can be a work-around having to build a big database of annotated images which is very costly especially in the medical field. We have also shown that out-of-the-box solutions are not well suited for this novelty imaging technique and there is limitless research potential since FFOCT is little explored in the domain of automatic diagnosis. Another idea that emerged during the evolution of this project, is that, in order to accurately assess the efficiency of such a model, we need to understand the reasoning learned by the machine. In our attempt to demystify deep learning, we looked under the hood of the network and visualized its "intelligence" under different forms: filters, activation maps, maximum activation input and offered a flavor towards the disambiguation of this black-box model.

The project opened many perspectives for future work, first of all, we will exploit the information of the abnormal tissue surrounding some cancerous zones by adding a new annotation class and establishing some relationship between classes; then we will try segmenting out the cancerous tissue taking as inspiration U-Net[34] and MiMo-Net[35] which are destined for biological

images. Moreover, Dynamic FFOCT broadens the possibilities of research even more, for example we can adopt composite architectures i.e. multiple networks, each taking as input a different type of data: FFOCT images, metabolic signal from D-FFOCT. We also intend to extend the multi-modal approach to clinical and genetic data.

These ambitious goals represent my research plan for the next 3 years and I am entirely grateful to my supervisor for giving me the chance to pursue a PhD that will allow me to continue on the path of this project which provoked my scientific curiosity.

References

[1] Christof Angermueller, Tanel Pärnamaa, Leopold Parts, Oliver Stegle. Deep learning for computational biology. Molecular Systems Biology Vol. 12. 2016.

[2] Yann LeCun, Yoshua Bengio, Geoffrey Hinton. Deep Learning Review. Nature, 2015.

[3] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 2017.

[4] Eugénie Dalimier, Denis Salomon. Full-field optical coherence tomography: a new technology for 3D high-resolution skin imaging. Dermatology, 2012.

[5] C Apelian, V Harms, O Thouvenin, AC Boccara. Dynamic full field optical coherence tomography: subcellular metabolic contrast revealed in tissues by interferometric signals temporal analysis. Biomedical Optics Express, 2016.

[6] Mehreen Adhi, Jay S. Duker. Optical Coherence Tomography – Current and Future Applications. Current opinion in ophthalmology 24.3, 2013.

[7] Freerk G. Venhuizen, Bram van Ginneken, Bart Liefers, Mark J.J.P. van Grinsven, Sascha Fauser, Carel Hoyng, Thomas Theelen, Clara I. Sánchez. Robust total retina thickness segmentation in optical coherence tomography images using convolutional neural networks. Biomed. Opt. Express 8, 2017.

[8] Leyuan Fang, David Cunefare, Chong Wang, Robyn H. Guymer, Shutao Li, Sina Farsiu. Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search. Biomed. Opt. Express 8, 2017.

[9] Cecilia S. Lee, Doug M. Baughman, Aaron Y. Lee. Deep Learning Is Effective for Classifying Normal versus Age-Related Macular Degeneration OCT Images, Ophthalmology Retina, Volume 1, Issue 4, 2017.

[10] Edward H. Shortliffe, A. Carlisle Scott, Miriam B. Bischoff, A. Bruce Campbell, William Van Melle, Charlotte D. Jacobs. 1981. ONCOCIN: an expert system for oncology protocol management. In Proceedings of the 7th international joint conference on Artificial intelligence - Volume 2 (IJCAI'81), San Francisco, CA, USA.

[11] Stanislav N. Tolkachjov et al. Understanding Mohs Micrographic Surgery: A Review and Practical Guide for the Nondermatologist. Mayo Clinic Proceedings, Volume 92, Issue 8. 2017.

[12] American Cancer Society. Cancer facts & figures 2016. Atlanta, American Cancer Society 2016. http://www.cancer.org/acs/groups/content/@research/documents/document/acspc-047079.pdf

[13] Sumaira Z. Aasi, David J. Leffell, Rossitza Z. Lazova. Atlas of Practical Mohs Histopathology. Springer. 2013.

[14] Li Fei-Fei, Rob Fergus, Antonio Torralba. Recognizing and Learning Object Categories, CVPR 2007 short course.

[15] Fabrice De Chaumont et al. Icy: an open bioimage informatics platform for extended reproducible research. Nature methods 9.7, 2012.

[16] Raphael Maree et al. Cytomine: An Open-Source Software For Collaborative Analysis Of Whole-Slide Images. Diagnostic Pathology, June 2016.

[17] Frank Rosenblatt. The Perceptron–a perceiving and recognizing automaton. Report 85-460-1, Cornell Aeronautical Laboratory. 1957.

[18] David G. Lowe. Object Recognition from Local Scale-Invariant Features. International Conference on Computer Vision (ICCV), 1999.

[19] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-Up Robust Features (SURF). Computer Vision and Image Understanding, 2008.

[20] Navneet Dalal, Bill Triggs. Histograms of Oriented Gradients for Human Detection. Conference on Computer Vision and Pattern Recognition (CVPR), 2005.

[21] Ivar Austvoll. Filter Banks, Wavelets, and Frames with Applications in Computer Vision and Image Processing (A Review). Scandinavian Conference on Image Analysis (SCIA). 2003.

[22] N. V. Chawla et al. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, vol. 16, pp.321–357, 2002.

[23] N. V. Chawla. Data Mining for Imbalanced Datasets: An Overview. Data Mining and Knowledge Discovery Handbook, Springer (pages 875–886). 2010.

[24] David E. Rumelhart, Geoffrey E. Hinton, Ronald J. Williams. Learning representations by back-propagating errors. In Neurocomputing: foundations of research. Cambridge, MA, USA. 1988.

[25] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems: 26th Annual Conference on Neural Information Processing Systems, 2012.

[26] D. Ciresan, A. Giusti, L.M. Gambardella, J. Schmidhuber. Mitosis Detection in Breast Cancer Histology Images using Deep Neural Networks. MICCAI, 2013.

[27] D. Ciresan, U. Meier, J. Masci, J. Schmidhuber. Multi Column Deep Neural Network for Traffic Sign Classification. invited, Neural Networks, 2012.

[28] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner. Gradient-Based Learning Applied to Document Recognition. Proceedings of the IEEE, 1998.

[29] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research, 2014.

[30] J. Masci, U. Meier, D. Ciresan, J. Schmidhuber. Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction. ICANN, 2011.

[31] Xavier Glorot, Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), 2010.

[32] Diederik P. Kingma, Jimmy Ba. Adam: A Method for Stochastic Optimization. ICLR, 2015.

[33] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, Ping Tak Peter Tang. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. ICLR. 2017.

[34] Olaf Ronneberger, Philipp Fischer, Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015

[35] Shan e Ahmed Raza, Linda Cheung, David Epstein, Stella Pelengaris, Michael Khan, Nasir Rajpoot. MIMO-Net: A multi-input multi-output convolutional neural network for cell segmentation in fluorescence microscopy images. ISBI, 2017.

Appendix

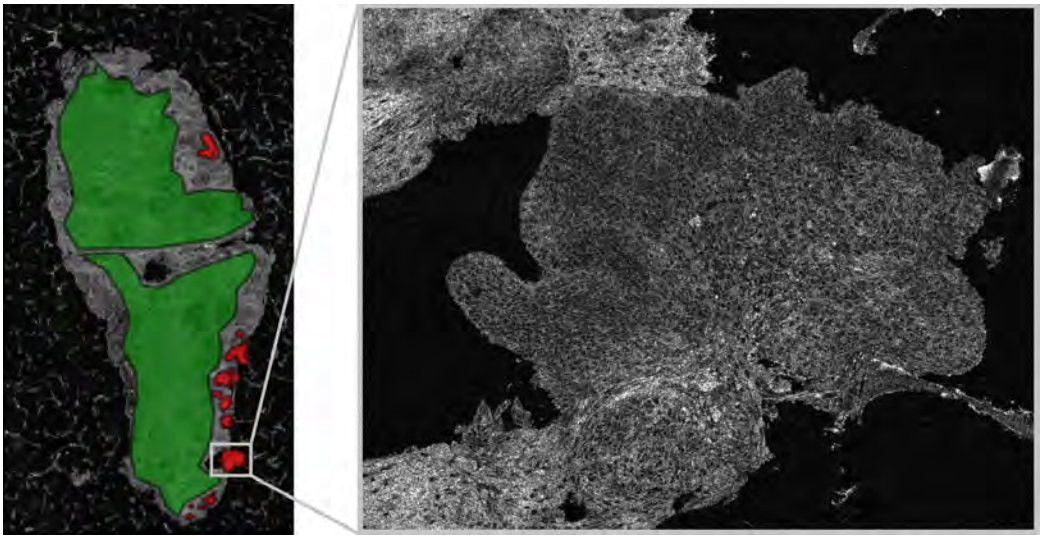
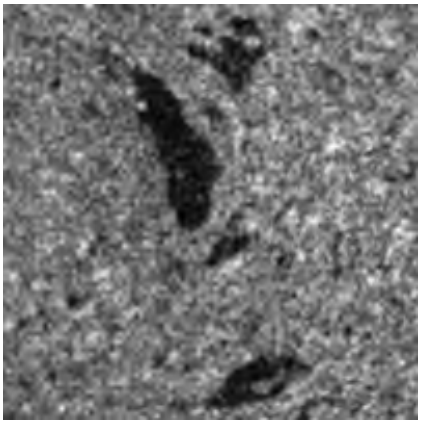
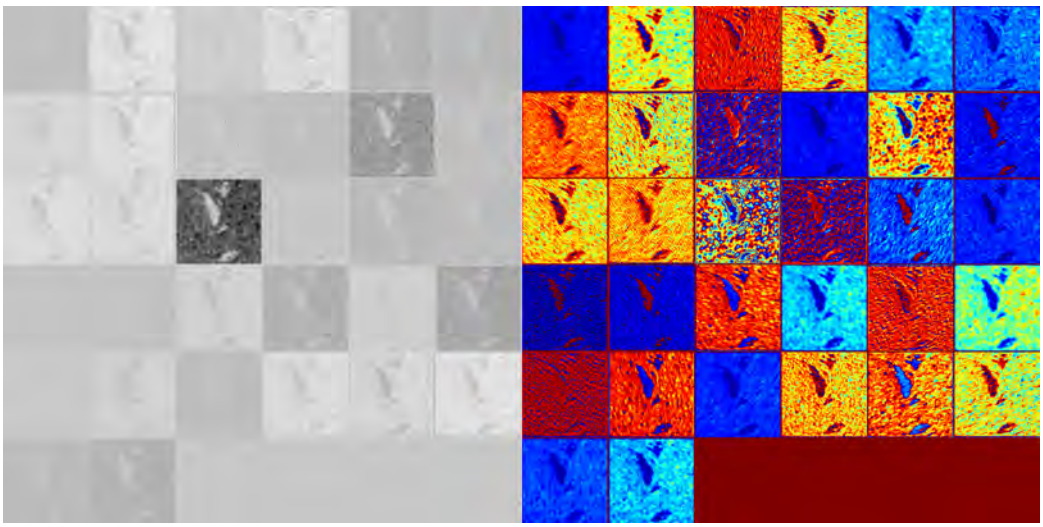


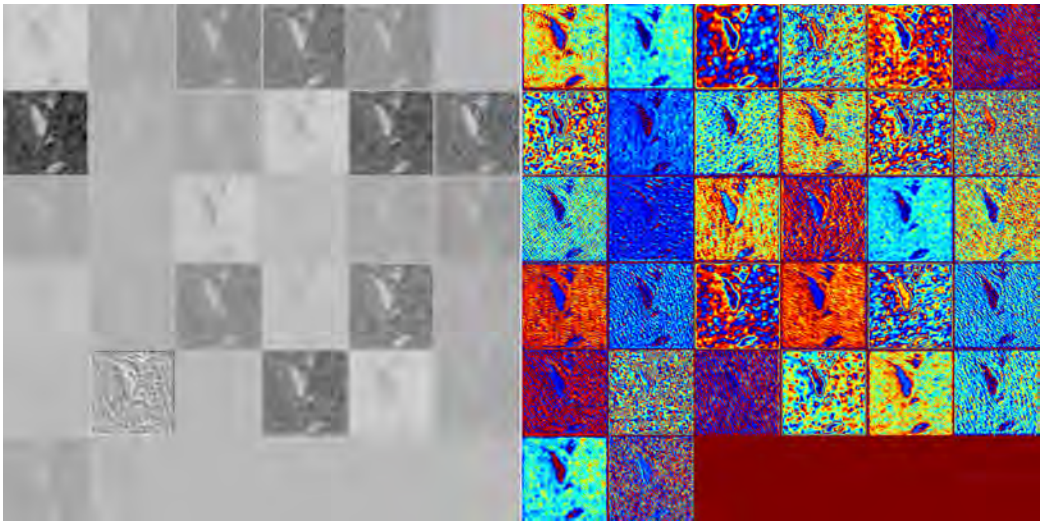
Figure 8: Annotated image and zoom on a **cancerous area**



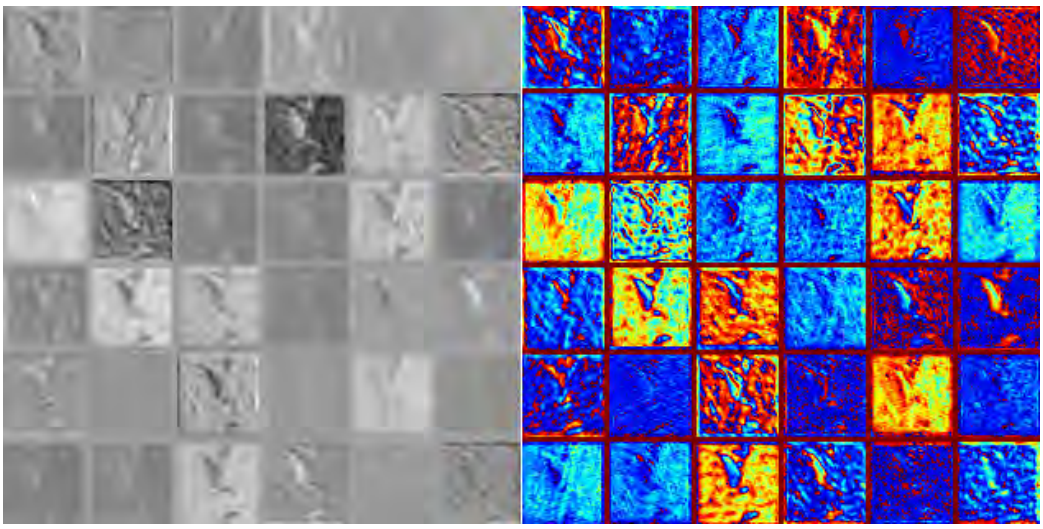
(a) Input patch (BCC)



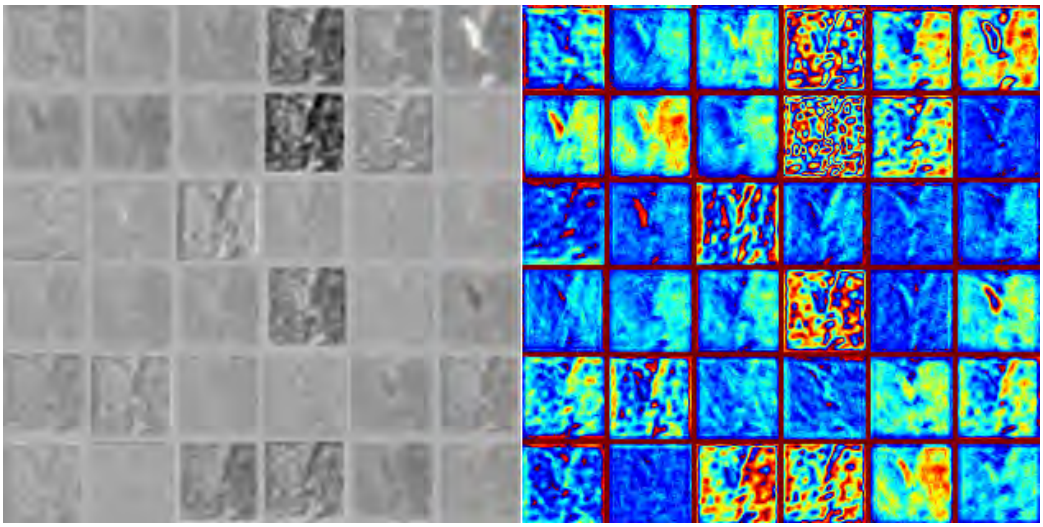
(b) Conv 1



(c) Conv 2

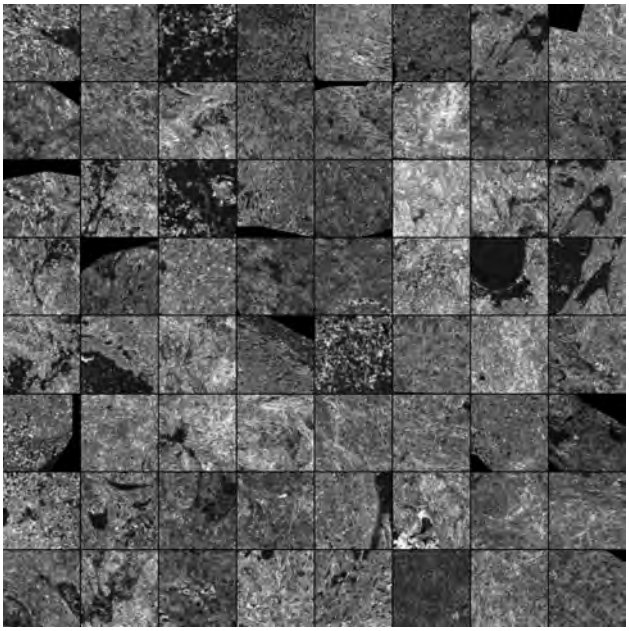


(d) Conv 3

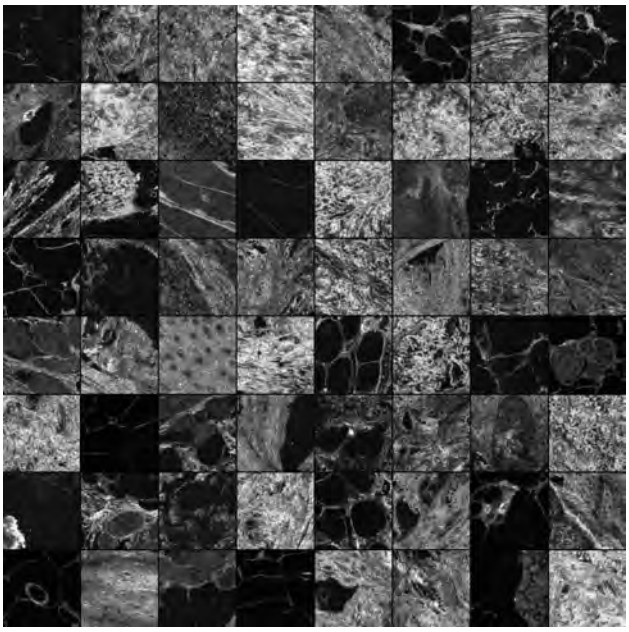


(e) Conv 4

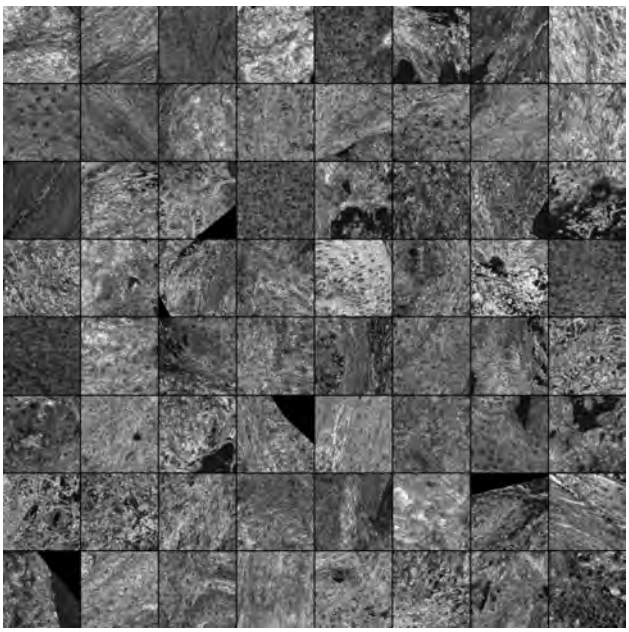
Figure 9: Input patch and **activation maps** corresponding to the filters in Fig.3 for the supervised architecture trained on 40 images (left: true values; right:with jet colormap showing high activation in red and low activation in blue)



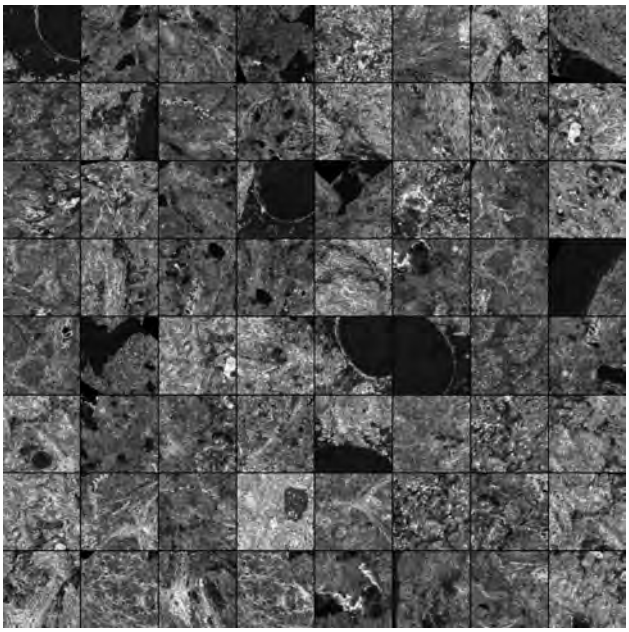
(a) True Positives (BCC)



(b) True Negatives (Normal)



(c) False Positives (Normal classified as BCC)



(d) False Negatives (BCC classified as Normal)

Figure 10: **Predictions** of the network trained supervised on 40 images