DETECTION AND CLASSIFICATION OF BREAST CANCER IN WHOLE SLIDE HISTOPATHOLOGY IMAGES USING DEEP CONVOLUTIONAL NETWORKS

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

MASTER OF SCIENCE

IN

COMPUTER ENGINEERING

By

Barış Geçer

July 2016

Detection and Classification of Breast Cancer in Whole Slide Histopathology Images Using Deep Convolutional Networks
By Barış Geçer
July 2016

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Selim Aksoy(Advisor)

Sinan Kalkan

Ramazan Gökberk Cinbiş

Approved for the Graduate School of Engineering and Science:

Levent Onural
Director of the Graduate School

ABSTRACT

DETECTION AND CLASSIFICATION OF BREAST CANCER IN WHOLE SLIDE HISTOPATHOLOGY IMAGES USING DEEP CONVOLUTIONAL NETWORKS

Barış Geçer
M.S. in Computer Engineering
Advisor: Selim Aksoy
July 2016

The most frequent non-skin cancer type is breast cancer which is also named one of the most deadliest diseases where early and accurate diagnosis is critical for recovery. Recent medical image processing researches have demonstrated promising results that may contribute to the analysis of biopsy images by enhancing the understanding or by revealing possible unhealthy tissues during diagnosis. However, these studies focused on well-annotated and -cropped patches, whereas a fully automated computer-aided diagnosis (CAD) system requires whole slide histopathology image (WSI) processing which is, in fact, enormous in size and, therefore, difficult to process with a reasonable computational power and time. Moreover, those whole slide biopsies consist of healthy, benign and cancerous tissues at various stages and thus, simultaneous detection and classification of diagnostically relevant regions are challenging.

We propose a complete CAD system for efficient localization and classification of regions of interest (ROI) in WSI by employing state-of-the-art deep learning techniques. The system is developed to resemble organized workflow of expert pathologists by means of progressive zooming into details, and it consists of two separate sequential steps: (1) detection of ROIs in WSI, (2) classification of the detected ROIs into five diagnostic classes. The novel saliency detection approach intends to mimic efficient search patterns of experts at multiple resolutions by training four separate deep networks with the samples extracted from the tracking records of pathologists' viewing of WSIs. The detected relevant regions are fed to the classification step that includes a deeper network that produces probability maps for classes, followed by a post-processing step for final diagnosis.

In the experiments with 240 WSI, the proposed saliency detection approach outperforms a state-of-the-art method by means of both efficiency and effectiveness, and the final classification of our complete system obtains slightly lower accuracy than the mean of 45 pathologists' performance. According to the McNemar's statistical tests, we cannot reject that the accuracies of 32 out of 45 pathologists are not different from the proposed system. At the end, we also provide visualizations of our deep model with several advanced techniques for better understanding of the learned features and the overall information captured by the network.

Keywords: deep learning, computer-aided diagnosis, whole-slide histopathology, saliency detection.

ÖZET

DERİN EVRİŞİMLİ AĞLAR İLE TÜM SLAYT HİSTOPATOLOJİSİ RESİMLERİNDE MEME KANSERİ TESBİTİ VE SINIFLANDIRILMASI

Barış Geçer
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: Selim Aksoy
Temmuz 2016

En ölümcül kanser tiplerinden olan meme kanseri, deri tipli olmayan kanserler arasında en sık görünen ikinci kanser tipidir. Meme kanserinde erken ve doğru teşhis tam tedavi için oldukça kritiktir. Son zamanlardaki tıbbi görüntü işleme araştırmaları bu konuda umut vadeden sonuçlar elde etmişlerdir. Bu sonuçların biyopsi görüntülerinin analizi sırasında daha doğru anlaşılması ve olası sağlıksız dokuların tesbitinde fayda sağlayabileceği düşünülmektedir. Ancak ne yazık ki, tamamen otomatik bir bilgisayar destekli teşhis için tüm slayt histopatoloji görüntülerinin işlenmesi gerekirken, bu araştırmalar genellikle özel olarak kesilmiş ve etiketlenmiş görüntüler üzerine olmaktadır. Bununla beraber tüm slayt görüntülerinin boyutlarının oldukça büyük olmasından dolayı kabul edilebilir bir işlem gücü ve zaman içerisinde işlenmesi güçleşmektedir ve farklı bölgelerde sağlıklı veya tümör dokularını farklı aşamalarda bulunması tüm slayt görüntülerde tümör tesbiti ve sınıflandırılmasını zorlaştırmaktadır.

Biz, ham tüm slayt görüntüsünden son teşhise kadar hızlı bir biçimde tümör tesbiti ve sınıflandırması yapabilen ve bunu yaparken güncel derin öğrenme tekniklerini etkili bir biçimde kullanan bir bilgisayar destekli teşhis sistemi tasarladık. Bu sistem uzman patologların sistematik çalışma akışı ve aşamalı detaya yakınlaşma tarzından esinlenerek geliştirilmiş olup temelde iki aşamadan oluşmaktadır: (1) teşhis için ilgili alanların tesbiti, (2) tesbit edilen alanların beş kanser tipine sınıflandırılması. Özgün ilgili alan tesbit yaklaşımımız uzmanların birden fazla çözünürlük seviyesinde verimli arama örüntüsünü taklit etmektedir. Bunun için dört adet derin ağ patologların tüm slayt inceleme kayıtlarından çıkartılan örneklerle eğitilmiştir. Daha sonra sadece tesbit edilen ilgili alanlar üzerinde daha derin bir ağ ve ardıl-işleme kullanılarak her bir tüm slayt görüntüsü

tek bir kanser tipine sınıflandırılmaktadır.

240 tüm slayt görüntüsü üzerinde yapmış olduğumuz deneylerimizde, tasarladığımız ilgili alan tesbit yaklaşımımız bu sorunun çözümünde en gelişkin diğer bir yöntemden daha verimli ve etkin çalıştığı gözlemlenmiştir. Bütün sistemin nihai sınıflandırması ise 45 patoloğun ortalama başarısının hemen altındadır. Ayrıca derin öğrenme öznitelikleri, farklı görselleştirme teknikleri kullanılarak incelenmiş ve öğrenilen bilgiler görüntülenmiştir.

Anahtar sözcükler: derin öğrenme, bilgisayar destekli teşhis, tüm slayt histopatolojisi, ilgili alan tesbiti.

Acknowledgement

Although writing this piece of appreciation is the finishing touch on my thesis, my heart will never forget those who provided me their invaluable supports throughout this period. I would like to express my profound gratitude to these unique personages.

It is difficult to overstate my gratitude to my supervisor, Selim Aksoy, who helped me to carry this heavy burden with his continuous motivation and encouragement at the course of this thesis. I kept going on the right path thanks to his sound guidance, immense technical knowledge and great efforts to explain things clearly.

My sincere thanks also go to Sinan Kalkan and Ramazan Gökberk Cinbiş, for their gentle support, constructive comments and thorough review of this thesis.

I also appreciate TUBITAK (Scientific and Technical Research Council of Turkey) for funding me as a student in an ARDEB project with the code 113E602.

I feel lucky for having an enjoyable atmosphere at the office with my wonderful friends: Onur, Ali, Can, Troya, Iman, Kadir, Nabil, Caner and others.

Life would be very difficult without the exceptional companionship and wise counseling of my very best friends: Furkan Can, Ikram, Zeki, Ali, Mustafa, Furkan Hopa. I feel fortunate and honored for having their valuable emotional support, genuine camaraderie, great entertainment, and compassionate caring whenever I need.

Feeling my family's inimitable support at my back have always encouraged me to persist no matter what. Although they barely understand what I am doing, their selflessness and self-sacrificing without hesitation for me have shown their sincere love. Without doubt, I am deeply indebted to my dear mother, father and sister. Thank you!

Most importantly, I cannot express enough gratitude to my beloved wife for her deepest love, warming smile, encouragement and quiet patience which are the secret ingredients that kept me going, especially in difficult times. Thank you for standing beside me throughout this work!

Above all, the most valuable contribution of this study to me is to contemplate the secret manifestations of Allah's divine names and attributes of which every scientific field relies upon. Certainly, I should say all praise be to Him, with the definition of *Bediuzzaman* in the *Letters* "*Since the perfections found in all beings which are the cause of acclaim and tribute are His, praise too belongs to Him. Acclaim and laudation, from whomever to whomever it has come and will come, from pre-eternity to post-eternity, all of it belongs to Him.*"

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The most widespread form of cancer among women is the breast cancer [3]. In the breast tissue, there can be many different types of deviations from a healthy structure, some being considered benign and some cancer. These deviations do not necessarily form a continuous spectrum of changes, and their detection and classification are not always straightforward for pathologists or CAD systems. Although people who live in the developed world have higher survival rates [4], patients have less chance with breast cancer in the developing countries. Medical image analysis promises to play an important role in helping experts in the analysis of histopathology images by improving the interpretation or indicating candidate disease locations [3]. Especially, computer aided diagnosis might be more beneficial in developing countries which are not lucky enough to have expert clinicians.

Traditional approaches to histopathological image classification involve supervised learning techniques that use manually selected regions of interest with class labels provided by pathologists. However, these methods are not directly applicable to the analysis of new generation whole slide images that contain multiple

areas with different levels of diagnostic importance. Thus, the identification and localization of diagnostically relevant regions of interest has emerged as an important initial step for whole slide image analysis.

Hand-crafted low- and mid-level features such as LAB, local binary patterns or graph related features have been widely used in medical image analysis in the past [5, 3, 6]. Although they are shown to be effective for discrimination of easier problems such as classification of healthy versus invasive cancerous regions, the information can be captured by those features is limited for more complicated tasks.

Those hand-crafted features are often used to train a classifier such as support vector machines or neural networks. Beside classifier learning, pattern classification methods that use feature learning approaches have been demonstrated to outperform those that use hand-crafted ones which have been studied over the years (i.e., SIFT, SURF, HOG etc.) [7]. One such approach is convolutional neural networks (CNNs) which were proposed by [8] for the first time in the early '90s. They have gained particular popularity in recent years thanks to increasing computational power by GPUs and availability of massive amounts of data on-line.

CNN's deep architecture enables to learn a hierarchy of features (i.e., flow in the order of pixel, edge, texton, motif, part, object [9]) from a given training data. Thanks to that hierarchy, it is successful on various computer vision tasks ranging from optical character recognition [10] to object recognition/detection [1], from scene labeling [11] to face recognition [12]. Besides, when the performances of CNN is investigated on very large data sets such as ImageNet [13] and compared to the human visual system, it is concluded that it outperforms humans at fine-grained recognition like animal species (i.e., 120 species of dogs in ImageNet) even in clear view [14].

Usually, after designing and training a CNN, it is limited to a particular input size, which might need to be varying for many vision tasks. Some studies suggest utilizing CNN to take arbitrary-sized inputs and generate a prediction

of corresponding resampled size. This kind of approach is first proposed by [15] and extensively investigated recently [16, 17]. This is called fully convolutional networks (FCNs) which process the whole image in one go for both feedforward computation and learning. Thanks to its dense convolutional nature, FCNs can be trained end-to-end efficiently and effectively which makes it suitable for detection and segmentation problems. An alternative to processing any-sized inputs is using sliding windows that does the computation for each window with traditional CNN. FCNs are advantageous over sliding windows by means of efficiency, since the windows contain overlapping regions that are processed multiple times whereas FCNs do the computation for one time for each pixel without repetition.

One disadvantage of CNN is being often considered as black box, which refers to hardness to understand of the inside prediction mechanism. While the recent CNN visualization studies show interesting facts about learned features, it might be interesting to see visualization of models learned from medical data for both medical and computer vision communities and may assist future studies to develop better CAD models.

1.2 Problem Definition and Our Contributions

In this thesis, we focus on both detection (localization) and classification of cancerous regions in whole slide breast histopathology images.

Regarding the detection task, we call diagnostically relevant regions of interest 'salient'. Main motivations of saliency detection can be grouped in three aspects: (1) In the literature, classification task is studied in a limited way on manually cropped cancerous regions whereas a fully automated CAD application necessitates computationally expensive whole slide image processing. ROI detection and classifying only the detected ROIs eliminates a significant amount of redundant computation and improves efficiency. An effective prior detection step should help to reduce false positives of classification without missing true positives as much as possible and improve the reliability. (2) Detection of relevant regions is

itself an important medical application which would lessen pathologists' workload significantly, where most of the cases are classified as health or benign [18]. Further, such system would also assure that no critical region is overlooked during diagnosis. (3) Since WSIs occupy enormous amounts of disk space, it is useful to extract regions that are more likely to be viewed in order to arrange priorities of viewing tools or compression applications.

Classification of detected regions is another significant task where a successful classifier would ensure a pathologist is not missing any dangerous possibility and might assist an inexperienced one during her diagnosis. Moreover, it would be a valuable tool in developing and underdeveloped countries without doubt.

Our contributions to address the aforementioned problems might be summarized as follows:

- We propose a saliency detection system by using advanced deep learning techniques, such that the machine will be taught to imitate actions of human pathologists for localization of diagnostically relevant regions in WSI. Saliency detectors are trained with the data extracted from screening records of experts such that a zoom action of a pathologist is used to construct one training sample.
- We study identification of five diagnostic categories of breast cancer in WSI by training a CNN. The network employed for this task is deeper and the data are greater compared to the detection part.
- For better understanding of the learned model, we visualize the resulting networks and observe which features play critical roles in differentiation of cancer categories.

1.3 Organization of the Thesis

In Chapter 2, we provide an overview of the related work about detection and classification in medical image analysis.

As background information for the upcoming chapters, basics of deep learning are explained in Chapter 3. After that, we demonstrate the state-of-the-art visualization methods of deep networks for better understanding of the learning process.

Chapter 4 describes the sophisticated methodology for extraction of the training set from pathologists' viewing records. Then, the FCN architecture of choice and the novel pipeline for efficient saliency detection in WSI are presented. After presenting the data set preparation and CNN design for classification of five diagnostic cancer classes, we explain the post-processing step for slide-based classification.

In Chapter 5, we provide experimental set-up, techniques used in the experiments, preparation of the ground truths labels and evaluation criteria for the detection and the classification tasks. We, then, present and discuss the performance comparison of our method with human experts and other methods. At the end, the visualizations of the learned features are illustrated with three different techniques.

Finally, we draw conclusion and elaborate on future works in Chapter 6.

Chapter 2

Related Work

There is a large body of work in the literature about medical image analysis based on low- and mid-level features [3, 5, 19]. While some of those are imported from fundamentals of computer vision such as texture [20], color, morphometric [21], topology and graph-based features [22, 23, 24, 25, 26], others are designed in accordance with clinical definition of pathologists [5] such as size and shape related features of objects, radiometric, densitometric, chromatin-specific [19], nuclei related features [27]. Moreover, some approaches include segmentation of various local structures like nuclei [27, 21] or gland segmentation [28] before extracting those features.

A significant amount of studies downgraded the problem into small patches cropped from whole-slide biopsies by hand. Even though this method gives an insight about which features can be useful for classification, they are not directly generalizable when the entire slide needs to be processed. While the simplest solution is dividing images into tiles and processing individually, this is, however, computational expensive on high resolution whole-slide images. An alternative method is extracting ROIs to be classified beforehand [29]. Some studies propose more efficient solutions such as detection in multi-resolution or multi-scale to reduce computational cost [30, 31, 32, 33, 34]. These approaches usually begin on a subsample of slides and increase resolution on interest regions until reaching

sufficient confidence. This methodology attempts to mimic the analyzing pattern of pathologists on whole-slides in a way. In particular, [34] presents more sophisticated pipeline for detection and classification of DCIS by utilizing superpixels for multi-scale progressive elimination. In contrast, [35] claims that low resolution might cause missing some details and proposes a greedy approach that randomly selects small patches to analyze whole-slide.

Although proposed methods are similar to our approach by means of mimicking pathologists' behavior that is going from lower resolutions to higher, we go beyond in that motivation and train our classifiers by their actual tracking data in order to detect exact interests of pathologists. Further, we consider saliency detection and classification of salient regions as two separate but sequential application which makes them modular and easy-to-use as distinct applications. Another disadvantage of mentioned studies is using the same hand crafted features in different levels of magnification which carry varied characteristics. This is handled by data-driven feature learning at each level in our method.

In medical image analysis, saliency detection on biopsy images is relatively unexplored problem. The studies that are dedicated to it [36, 2], contain disadvantages of hand-crafted features and running on a single resolution. Although [36] presents perfect results, we suspect that results are overly optimistic since evaluation is done on cropped patches instead of whole-slides. [2] measures the performance on whole-slides for the same data set that we use for evaluation, thus, in Chapter 5, we compare our results with it for saliency detection.

The use of existing hand-crafted features or designing new ones require expert knowledge about the field. Moreover, it is common that a combination of such feature descriptors is required to extract the most useful information in local patches. Recent studies [7] show that, in many computer vision tasks, the classic object detection/classification pipeline that uses hand-crafted features are less successful than feature learning approaches (i.e., CNN, autoencoders) which does not need expert knowledge.

Feature learning approaches are heavily applied to medical image analysis domain as well. In study [37], authors compare learned features by CNN with numerous state-of-the-art hand-crafted features in the literature at recognition of invasive ductal carcinoma tissue regions in WSI and show that learned features outperform hand-crafted ones. Some abilities of the deep learning are also applied to the field. For example, weak learning methods such as multi-instance learning are employed in medical images by [38, 39] and observed slight improvement. In [40], multi-scale CNN approaches [11] are adopted to be effective in capturing both textural and abstract information for segmentation of cervical cytoplasm. Transfer learning is first applied by [41] that adapts features trained with breast histopathology images to medulloblastoma tumor differentiation. Having that features learned from breast histopathology performed better than features learned from natural images of ImageNet data set, suggests that features with better characteristic of the target data and the task should achieve better performance. Fully convolutional networks are used by [42] for efficient detection after thresholding non-salient regions. Different network architectures, which yield the best performances in object detection challenges, are examined in medical image analysis by [39].

Not long ago, [43] won the Camelyon Grand Challenge 2016 where it achieved better performance than hand-crafted features and slightly lower performance than human accuracy. Moreover, it improved human performance significantly by combining prediction probabilities of the two. But the running time of this approach is not scalable for a practical application for clinics as it processes WSIs at 40X magnification.

Our study presents classification of five diagnostic classes of breast cancer and visualization of abstract level of information captured by networks. To the best of our knowledge, no study is done for classification of multiple classes of cancers, most of them focus on binary classifications such as cancer vs. healthy or detection of a class (i.e., DCIS, INV) which are relatively simpler problems. Furthermore, other than first layer reconstructions, we did not come across with any advanced deep learning visualization technique applied to the domain.

Chapter 3

Background on Deep Learning

This chapter briefly outlines the basics of deep convolutional neural networks and explains the visualization approaches used for better understanding of the learned models.

3.1 Basics

The goal of deep learning is to build a data-driven solution to model a particular problem with sequence of layers that are developing from low-level features (i.e., edges, T junctions) to more abstract representations (i.e., human head, keyboard). CNN is a visual form of hierarchical networks where discriminative, representative features are learned (not the design of the network which is handcrafted) for end-to-end prediction from raw pixels of the original image to final scores. Each layer consists of a differentiable function that inputs a 3D volume to be transformed into an output 3D volume, such as linear filters, non-linear activation functions, subsampling operators and an objective function. An optimization algorithm, such as gradient decent, runs iteratively and updates parameters of those functions in such a way that it takes one step (learning rate) in the direction of fastest descent of the loss and, ultimately, reaches a local optima in the

objective function space. Although it is rare to achieve global optima, most of local optima points are good enough and quite close to global optima. We now explain some of the fundamental terms about CNN below:

1. *Fully-connected (FC) layer:* This layer, as in traditional multilayer neural network, consists of weighted connections from all units of the preceding layer to all units of the subsequent layer.

2. *Convolutional (CONV) layer:* For visual learning, it is possible to make full neural connections from all to all, the computation time, however, would not be scalable for proper training as the number of parameters to be trained increases exponentially. This problem is overcome with the assumption of that input images possess similar characteristic in any spatial position. That is to say, for example, possibility of having a horizontal edge in the middle of the image is the same as having it at any other position. This implies that features learned from the middle part of the images are also useful in any other part. Therefore, it is sensible to make local shared connections whose weights are kept the same for every spatial position of the image.

   Convolution is an operation where a kernel (or filter) slides (or is convolved) over a vector or matrix and the dot product of corresponding local region and weights of kernel computes the corresponding output. This operation is widely used in image processing for detection of desired patterns and mostly done in 2D.

   Shared neural connection solution can be applied by a convolutional layer which carries out the core function of CNN approaches. They consist of multiple 3D convolution operations which are applied to volumes coming from the preceding layer and produce the output volume which has the same width and height as the input. Each convolution kernel is a volume that has a relatively small width, height (i.e., 3, 7, 11) and depth same as input volume. The number of kernels applied determines the depth of the output volume where each depth column corresponds to the response of the input volume to a filter that is selective for a particular pattern. Trough learning algorithms, the network will learn filters and they, eventually, converge to

discriminative features for the training data.

Although there is no right number and size for kernels, as they are design parameters and are problem-oriented, better models can be designed with broad experiments [1] or by visual analysis [44]. The size of the region a kernel is operated on the original image, is dependent on kernel's size and place in the network, and this is called receptive field. Receptive fields of kernels are expected to be enlarging over the layers and should be big enough at the end to capture significant information. In order to increase spatial tolerance and, in some cases, make the network shallower, filters can be slid with a stride more than one which can be viewed as down-sampling the output of convolution as well.

3. *Deconvolutional layer:* Deconvolution is the reverse operation of convolution which may be considered as learnable upsampling instead of being fixed (i.e., bilinear) in a network. When it is integrated in a backpropagation algorithm, it provides end-to-end learning from dense label maps with pixel-level precision which is useful for detection and segmentation tasks. We refer readers to [16] for details about deconvolutional layer.

4. *Non-linearity:* Most of the time, the problem that is tried to be modeled by CNN is not linearly separable. Thus, non-linearity should be introduced to the network in order to solve such problems. In other words, without non-linear activation functions, a network's output would just be a linear combination of the input. In the recent studies, it is proven that the rectified linear unit (ReLU) (i.e., thresholding at zero $f(x) = max(x, 0)$) as a non-linear activation function is the most efficient and effective compared to other functions (i.e., $tanh$) [45]. In CNN architectures, a convolutional layer is mostly followed by a non-linearity gate.

5. *Max pooling layer:* Max pooling is a method for taking the highest activation unit into account in a given interest region where ignoring other units, and therefore is a down-sampling operation which actually reduces the number of parameters and, thus, computation time. This operation is mostly done after convolution and non-linearity operations and gives the network tolerance to translation of interest pattern in the input.

6. *Dropout:* Overfitting is one of the most challenging and common problem of machine learning and particularly deep learning. The dropout approach, which is proposed by [46], deals with this problem by simply disabling some of the units during training (i.e., 50%) randomly in each epoch and this method is usually applied to FC layers.

7. *Objective function:* In other name, loss function (i.e., softmaxloss) defines the optimal solution such that the gradient descent algorithm will update weights toward the negative direction of its gradient and eventually converges to a local optima.

8. *Gradient based optimization:* After initializing all weighted connections randomly with zero mean and small variance, they are optimized according to given data and the objective function with conjunction of gradient descent and back-propagation algorithms.

   First, all of or a subset of training data is fed to the network forwardly. Then the error is estimated from the difference between the prediction and the expected output. Secondly, all the weights in the network are updated in the negative direction of the gradient derivative of the objective function by calculating the derivatives with back-propagation algorithm. Since it is cost-inefficient to run this algorithm for all the data for each update, mostly the stochastic gradient descent (SGD) algorithm is used in practice which divides training data into subsets and run gradient descent algorithm for each. Although this is a greedy approximation of the gradient descent algorithm, it converges to an adequate optima point in reasonably less time.

9. *Fully-convolutional Network:* Traditional single convolutional networks with sliding windows suffer from two main problems in detection/segmentation tasks: (1) fixed input size limits possible applications and it is not clear how sliding windows should be managed (i.e., step size), (2) unnecessary computation of overlapping regions multiple times decreases the efficiency drastically.

   Fully convolutional networks (FCN) are generalized version of CNN which accept any-sized inputs and produce a dense output map in respective size.

Instead of scanning whole image with classic CNN which cause repeating computations of coinciding regions, this approach is more efficient by means of learning and application time, where images are processed as a whole without redundant computation.

In order to alter a CNN to FCN, fully-connected layers should be convolutionalized (as, in fact, they are convolutions where input and convolution dimensions are equal) and objective function should be remodeled according to new spatial output map.

3.2 Visualization

Over the years, CNNs have been considered as black box and improving a network's performance regarded as only manipulating the hyper-parameters blindly. Recent studies enlightened this puzzle and developed novel visualization methods for better understanding of the inner units of a network. Interpretation of a learned model is important in order to spot ways of improvement or locate the obstacles. For some, this progress was even more exciting by means of discovering that the learned features are the ones that are attempted to be designed by hand for decades. As a result, the initial goal that is the hierarchical learning of features in a logical flow (i.e., edge, texton, motif, part, object) was demonstrated to be achieved more or less by CNN.

Therefore, we apply several visualization techniques to our models in order to understand which features are important for identification and discrimination of cancer types. Observing the information captured by our network might help to improve it or to give intuition to design new ones in the future studies. What is more is that discovering new features that are regarded as discriminative by our network, might be interesting to pathologists as this might help to improve their understanding and performance during the diagnosis which is still an open problem. Additionally, visualization of different patterns that are diagnostically relevant can be beneficial during the medical education.

CNN visualization techniques can be classified into three main categories as done by [47]:

1. *Occlusion:* This trivial approach is first proposed by [44] and improved by [48]. It is all about simply considering the network as an unknown function and measuring the influence of input pixels. The method first measures the activation of an image for a particular class, then starts distorting the input image within small windows by replacing each window with television noise independently. While distorting each window, it monitors the differences in the activation and produces a heatmap that shows how much each particular window changes the likelihood of the class either in positive or negative direction.

2. *Deconvolution:* This method is proposed by [44] and [49] at first and improved by [50]. The goal is to produce the visualization of the contributions of pixels of an image by tracking the activations during the feedforward pass and deconvolving toward backward direction until the input.

3. *Activation Maximization:* [51] proposed a method to synthesize an image that maximize the activation of a neuron. This image resembles the pattern which the neuron is sensitive the most. Since this approach led to unnatural images, [49] suggested to add a L_2 norm regularization term to optimization objective. This term encourages pixel values staying in the rage of natural images (which is usually [0, 255] or [-128, 128]). In addition to that, [52] includes total variation regularizer which smooths out the spikes caused by max pooling layers during the reconstruction. They also generalized L_2 norm regularization to L_p norm and use larger values of p (i.e., $p = 6$) that results in better visualizations. [53] adds a non-parametric path prior regularization on top of others, but we do not include it in our experiments. The drawback of this approach is that balancing between the loss and the regularization terms requires some attention, and otherwise it leads to unrecognizable reconstructions for human observers.

Chapter 4

Methodology

This chapter focuses on the novel methodology that we propose for detection and classification in whole slide breast histopathology images by employing deep CNN models. We consider these two tasks as sequential yet distinct applications.

Section 4.1 describes the proposed procedure that uses advanced deep learning techniques to mimic the actions of pathologists for the detection of regions of interest in WSI of breast biopsies. First, training data are gathered from the tracking records of pathologists' viewing behavior while they were interpreting whole slide images (Section 4.1.1). The actions that correspond to zoom events and panning motions are identified by using a set of rules that results in candidate salient regions at different magnification levels. Section 4.1.2 explains how these regions are used to train fully convolutional networks[1] that are particularly efficient for detection tasks in arbitrarily sized images. Finally, four separate networks that model the pathologists' screenings at different magnifications are combined in a sequential pipeline for effective and efficient processing of large images where areas that are found to be non-salient are incrementally eliminated from lower to higher resolutions as shown in Section 4.1.3. The resulting probability maps quantify the diagnostic relevance of all pixels in whole slide images.

[1] Although these networks are also CNN, we prefer to call them FCN throughout the text in order to avoid confusion.

In Section 4.2, we explore the performance of the learned hierarchical features on identification of several diagnostic breast cancer classes. We first prepare a training set by using only the most informative regions marked by experts in Section 4.2.1. Then, it is introduced to one deeper CNN compared to FCN design to be trained for classification of five cancer categories. At the end, we explain a post-processing step to obtain slide-based predictions from pixel-based probabilities.

4.1 ROI Detection

4.1.1 Data Set Preparation

We use a data set of 240 digital breast histopathology images that are collected as part of the digiPATH project [2]. The H&E stained biopsy slides were scanned at 40X magnification in average size of 100,000×80,000 pixels and were labeled by 196 pathologists where 3 of them are world-class experts in the field. Several studies [54, 55] show the huge disagreement among those individual pathologists' interpretations of the slides. For ease of use and efficiency, images are six times subsampled by a factor of 2 to construct a spatial pyramid that resulted in 20X, 10X, 5X, 2.5X, 1.25X, 0.625X magnification versions of the original image and divided into 180 training and 60 test images.

During diagnosis of high-resolution pathology images, pathologists use software that allows zoom in/out and panning actions. The software works similar to the web mapping applications (i.e., Google Maps) where it shifts between the mentioned resolutions when the user changes the zoom level. If requested zoom level is not one of the sub-sampled resolutions, then the software subsample-on-time from the closest higher resolution. The rectangular part of the image that is visible on the pathologist's screen (viewport) is tracked while she navigates over a range of magnifications and regions for diagnosis. At the end of the viewing session, the pathologist also marks a bounding box to indicate a sample region

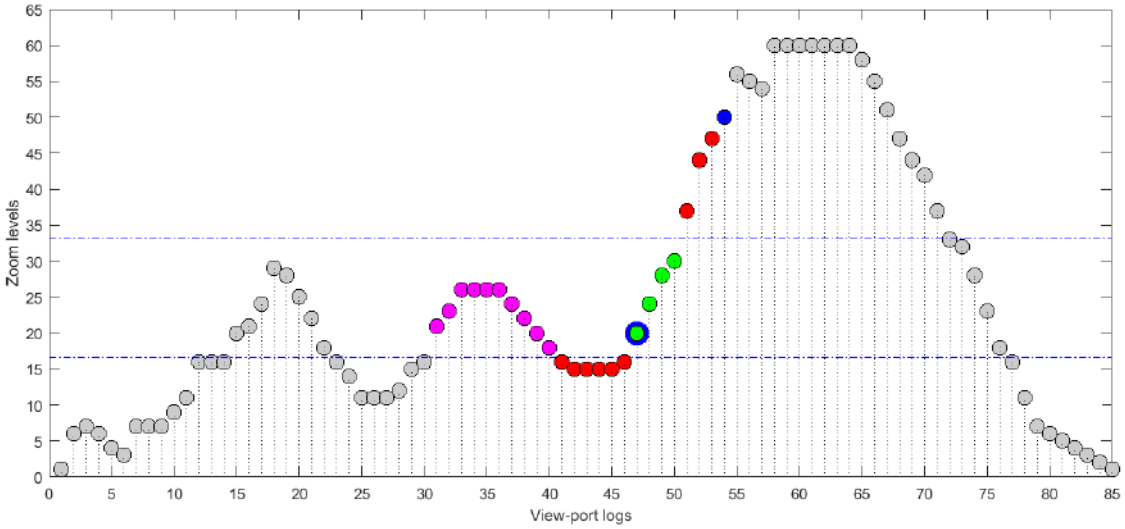


Figure 4.1: Zoom levels of a cross-section of view-port logs. Extraction of a zoom action is shown where blue dot represents the inner window's log, horizontal blue dashed lines indicate the range of possible outer window candidates' zoom level due to Equation (4.2), the red dots show logs eliminated because of this limitation, pink dots depict logs that violate Equation (4.3). Green dots satisfy all the conditions and the earliest one is considered as the outer window of the zoom action extracted. Every log in the set is considered as the inner window and this process is repeated. *Best viewed in color.*

for diagnosis. The view-port logs are recorded 4 times in a second and each log includes coordinates of the visible part, the zoom level and the time stamp information. Let us call view-port logs l_t for each expert's analysis on each image, where $t = 1, 2, \ldots, T$ stands for time series.

A well-prepared training set is one keystone for effective training of a deep network. In such complicated data sets, presenting the data to the network plays a critical role; therefore, here we are mostly focused on this part. We propose the following procedure where we are motivated to spot experts' zoom actions. Intuitively, we presume that every visited window is considered as salient by the expert at some previous window. In this fashion, those windows should be neither too close nor too far to the visited window by means of zoom level and there should not be any zoom-out action between the two.

First, we define a set of conditions that assure that all zoom actions are explored accurately such that an inner window appeared to be salient to the expert when she was looking at the corresponding outer window. Then, we consider all the log records in our set as the zoomed window (l_j) and the candidate preceding logs as their outer windows (l_i), and collect log pairs (l_j, l_i) as follows:
Find smallest i such that

$$i < j \tag{4.1}$$

$$zoom(l_j)/3 \leq zoom(l_i) \leq zoom(l_j)/1.5 \tag{4.2}$$

$$zoom(l_k) > zoom(l_i), \quad \forall k \in \{i, .., j\} \tag{4.3}$$

An example is illustrated in Figure 4.1 where outer window (l_i) of the zoomed window $l_{j=54}$ (blue dot) is being discovered. We filter out the preceding logs that lies out of the zoom level range defined in Eq.(4.2) such as red dots. Pink dots are eliminated for having zoom-out actions between them and $l_{j=54}$ (i.e., logs with lower zoom level such as red dots). Finally, among the remaining candidate logs (green dots), the earliest log is chosen to be the outer window (the one surrounded with blue).

After discovering all the zoom actions, pairs that contain common outer windows (l_i) are grouped and each group is used to create one data sample where the input is raw image corresponding to the common window l_i and the label is a same sized binary map where the union of l_js' windows in the group is positive (salient). An example to this mechanism can be seen in Figure 4.2 where outer window is shown with red dot (i.e., $l_{i=47}$) and the group of corresponding zoomed windows are circled with red (i.e., $l_{j=54...67}$) which are then used to construct the label.

For every image, tracking logs of each of the three experts who are the only ones having viewing records for all images, are processed individually by the above algorithm and union of the all extracted samples formed the training data. Yet, there are 4 detectors to be trained, thus the data need to be split into 4 subsets. Since the detectors are hierarchical, the split is done according to zoom levels of the outer windows. Therefore, we determine 4 ranges of $zoom(l_i)$ values as shown

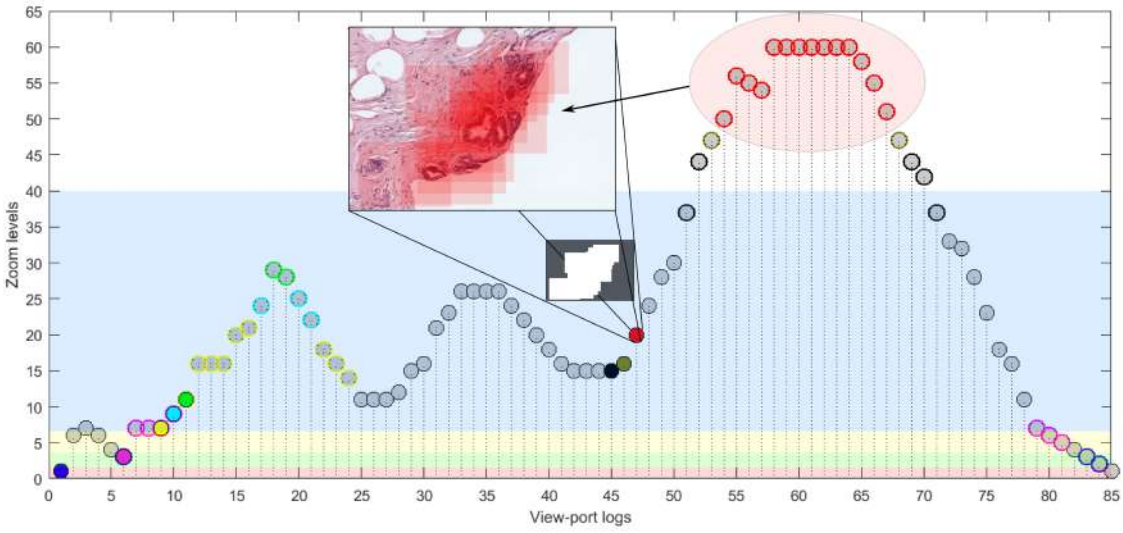


Figure 4.2: The same cross-section as in Figure 4.1. Groups of zoom action pairs are painted to the same color where if the color fills the blob, it is the outer window, if the color surrounds the blob, it is one of the zoomed windows of the group. An example of data sample construction is shown for the red group. Ranges of data split are shown with different colors; (red): 1. detector ($1 \leq l_i \leq 1$), (green): 2. detector ($2 \leq l_i \leq 3$), (yellow): 3. detector ($4 \leq l_i \leq 6$), (blue): 4. detector ($7 \leq l_i \leq 40$). *Best viewed in color.*

in Table 4.1 and Figure 4.2 with different colors. For example, for detector 3, we consider zoom action pairs that have l_i whose zoom level is lower than or equal to 6 and higher than or equal to 4.

At the end, we form the four training sets that consist of a total of 66,144 images with 535 $\times$ 416 pixels average size. Total number of pixels labeled as negative is around 5 times as many as those which are labeled as positive. Other statistics of the training set concerning subsets are shown in Table 4.1.

Table 4.1: Statistical information about the training data.

Detectors	$zoom(l_i)$ **ranges**	**Magnification**	**Number of Samples**	**Average Width**	**Average Height**	**Number of Positive Pixels**	**Number of Negative Pixels**
FCN-1	1-1	0.625X	40,209	564	430	77,031,781	717,253,531
FCN-2	2-3	1.25X	7,297	491	410	35,702,296	81,060,840
FCN-3	4-6	2.5X	9,125	498	392	48,828,834	92,800,606
FCN-4	7-40	5X	9,513	483	388	40,590,231	100,196,457
Total	-	-	66,144	535	416	202,153,142	991,311,434

4.1.2 A Deep Architecture for Saliency Detection

Another keystone for effective feature learning is network configuration which includes network architecture and other learning parameters (i.e., learning constant, momentum). Drawing a network design from scratch is a challenging task and requires extensive domain expertise. As discussed, we do not deeply investigate deep learning in this aspect but rather the way of applying well-studied deep-net solutions to our unexplored problem. Therefore, for all parameter and network choices, we draw inspiration from VGG network [1] because of its recent success in ImageNet challenge [13] and simpler deep-net models. Nevertheless, we assure that the sizes of the receptive fields of the CONV layers fit to fundamental elements of biopsies such as nuclei, ducts, lumen or a particular tissue pattern as seen in Figure 4.4.

The input of our networks are arbitrary sized ($m \times n$) RGB images that are collected as explained in Section 4.1.1. Input image is preprocessed by subtracting the overall mean of RGB values of training set images from each pixel. The image is then passed through 3 similar convolutional layers, as in [1], where filters have a very small width and height (3×3) followed by a ReLU non-linearity unit. Convolutional stride and spatial padding is set to 1 pixel such that the spatial resolution is preserved. ReLU is followed by the max pooling operation with a 3×3 pixel window and a stride of 3 after the first layer and a 2×2 window and a stride of 2 after other layers. This three convolutional layers are followed by another convolutional layer with 4×4 window size and convolutional stride of 4. This layer includes a ReLU non-linearity but there is no max pooling operation. After that there is one fully connected layer (which is, in fact, a 1×1 convolutional layer in FCN case) followed by dropout operation with 0.5 rate. The network continues with a deconvolutional layer with upsample rate of 16 times and cropping 32 pixels from all sides. Number of filters in all layers are $32, 32, 64, 128, 2$ respectively.

Size of the resulting map is relative to input size ($m/3 \times n/3$) as an advantage of fully convolutional design which improves precision of segmentation. The final

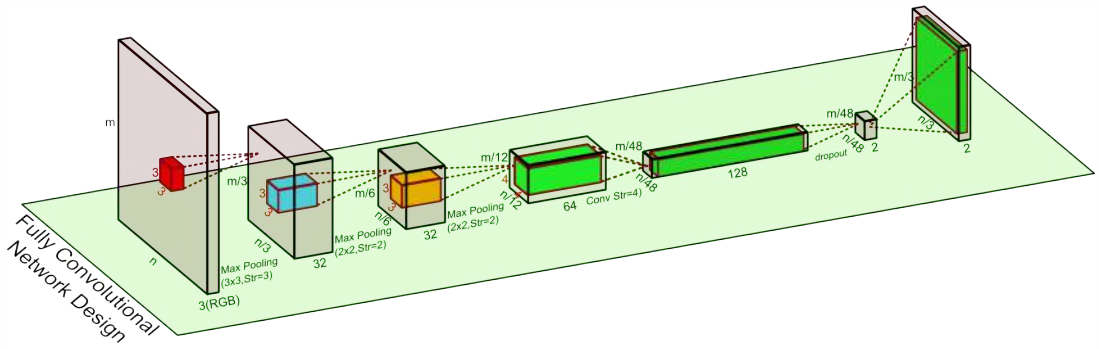


Figure 4.3: Designs of the Fully Convolutional Networks (FCNs). Hierarchical visual representations are learned with the simplistic design choice which is inspired by [1]. All the convolution layers are followed by a ReLU normalization layer. Note the deconvolutional layer at the end.

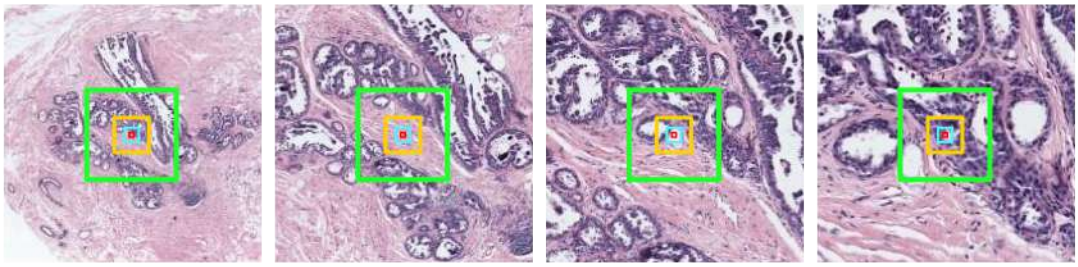


Figure 4.4: Receptive fields of CONV layers shown in Figure 4.3 overlaid on 0.625X(a), 1.25X(b), 2.5X(c), 5X(d) magnification images. This figure gives us intuition about what kind of information is captured by each layer of our network. Colors are selected according to their corresponding layer in Figure 4.3.

layer is connected to 'softmaxlogloss' objective function layer while training but after training we remove that layer and add 'softmax' layer to estimate class probabilities which are between 0 and 1. Graphical representation of this architecture can be seen in Fig.4.3.

4.1.3 Pipeline

We propose a pipeline that gradually eliminates insignificant regions efficiently in four successive steps where the ultimate output is a saliency map of the input image. A given image is processed by the four networks that are trained to handle

images from different resolutions. For example, let us call the input image with Φ which has 40X resolution. First, the spatial pyramid of Φ is extracted where, in fact, at this state of the framework, we only use 0.625X, 1.25X, 2.5X, 5X resolution levels which are called Φ_1, Φ_2, Φ_3, Φ_4 respectively. We, then feedforward Φ_1 to the first network (FCN_1) to produce a saliency map Θ_1 where the regions above a certain threshold are fed to the second network in two times higher resolutions (Φ_2). We repeat the same procedure a total of four times. Then we compute the weighted geometric mean of thresholded output of four networks denoted by $\Theta_{1,2,3,4}$ by the following formula:

$$\Theta = \prod_{i=1}^{4} \Theta_i^{\omega_i / \sum_{i=1}^{4} \omega_i}, \tag{4.4}$$

where $\omega_i = (\frac{1}{2})^{4-i}$. The output maps are thresholded in such a way that pixels below the threshold are set to the minimum value of the pixels above the threshold. This is in order to not to lose the saliency information obtained by earlier FCNs, while preserving the order of pixel values (i.e., the pixels below the threshold cannot have higher values than those above it in the geometric mean). The formulation of the thresholding stage is as follows:

$$\Theta_i(x,y) = \begin{cases} FCN_i(\Phi_i(x,y)), & \text{if } (x,y) \in \Omega_i \\ \min\limits_{\forall (x',y') \in \Omega_i} FCN_i(\Phi_i(x',y')), & \text{otherwise} \end{cases} \tag{4.5}$$

$$\Omega_i = \{(x,y) \in |\Theta_{i-1}|_\tau\}, \quad \text{for } i > 1 \tag{4.6}$$

where Ω_i denotes a set of pixels which is initially all the pixels in the image ($\Omega_1 = \{(x,y) \in \Phi\}$) and $|.|\tau$ denotes thresholding Θ_i adaptively such that lowest τ percentage of values of Θ_i is removed from the set of pixels to be processed in the subsequent steps. Tuning the parameter τ will be discussed in Chapter 5. Note that all Θ_i maps are scaled to same resolution and geometric mean is done pixel-wise.

This design is motivated to resemble the efficient search pattern of expert pathologists where they detect diagnostically relevant ROIs at a given resolution than zoom in to examine the region at better resolution. They repeat this procedure multiple times until they reach the ultimate regions of interest at highest

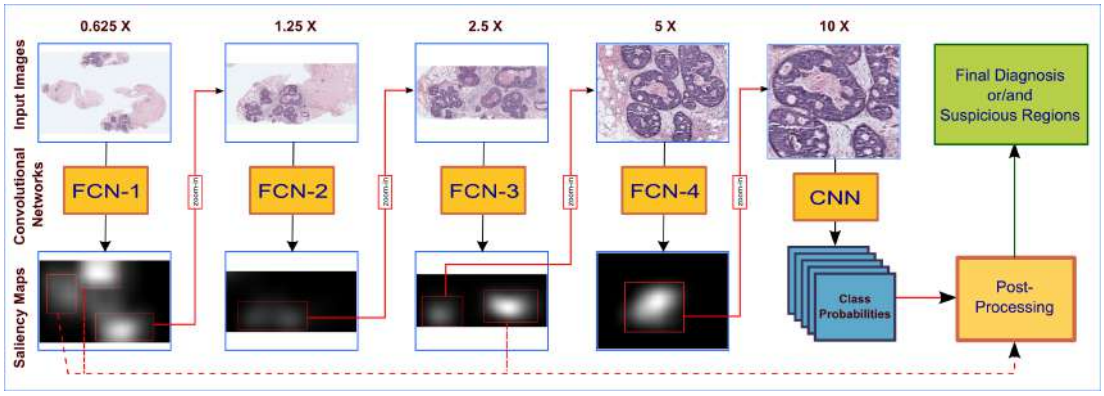


Figure 4.5: Overview of the proposed framework. Salient regions are detected from a WSI by feed-forward processing of FCN-1. Each connected component above a threshold is zoomed-in on the input image and processed by FCN-2. This process is repeated four times and the detected salient regions are processed by the classification CNN to obtain probability maps of five diagnostic classes. Then classifications of all salient regions are combined by post-processing to determine final diagnosis. With this hierarchical procedure, the goals are: (1) improving efficiency, (2) capturing information in all the same zoom levels as pathologists perform.

resolution and as a result, they avoid superfluous efforts [56]. Similarly, with the presented progressive procedure, we eliminate redundant computation, i.e., the whole image would be processed in 5X resolution with 63 times more computation compared to 0.625X resolution.

4.2 ROI Classification

4.2.1 Data Set Preparation

We will use the digiPATH data set introduced in Section 4.1.1, which is also suitable for classification as images are labeled into five different diagnostic categories of breast cancer (i.e., Non-proliferative changes only (NP), Proliferative changes (P), Atypical ductal hyperplasia (ADH), Ductal carcinoma in situ (DCIS), Invasive cancer (INV)). Similarly, only the same 180 images are used for training and 60 for testing in 10X magnification by keeping class distribution ratios the

equivalent as much as possible.

Whole slide biopsies are huge images and, thus, only a tiny part can change the decision of a pathologist to a cancer type even though remaining part consist of perfectly healthy tissue. Moreover, one WSI might contain multiple categories of cancerous tissues which may confuse both medical experts and computer systems. Therefore, expert pathologists are asked to draw boundaries of the most representative area for the most critical diagnosis found in that whole slide in consensus meetings. We, then sample patches size of 100×100 pixels from the training set where at least 80% of a patch lies within the union of those boundaries. Sampling is done with 50 pixel stride which means neighboring samples contain 50×100 or 100×50 overlapping regions. The resulting data set consists of 1,272,455 images distributed to five classes as shown in Table 4.2. The reason why we did not fully-convolutionalize the classification CNN for the training instead of patch extraction is that this CNN is deeper and contains more than 30 times more parameters compared to the detection FCN, therefore training fully-convolutionalized version of classification CNN would take around 4 months whereas training one detection FCN takes about 3 days and training the classification CNN takes about 10 days.

Table 4.2: Distribution of the training data of classification task

Category	**Number of Samples**	**Percentage**
NP	86,397	6.79%
P	242,434	19.05%
ADH	92,231	7.25%
DCIS	371,385	29.19%
INV	480,008	37.72%
Total	1,272,455	100.00%

4.2.2 A Deep Architecture for Classification

Classification of five cancer categories is more challenging than saliency detection, thus, requires deeper network design and further training data. The prepared

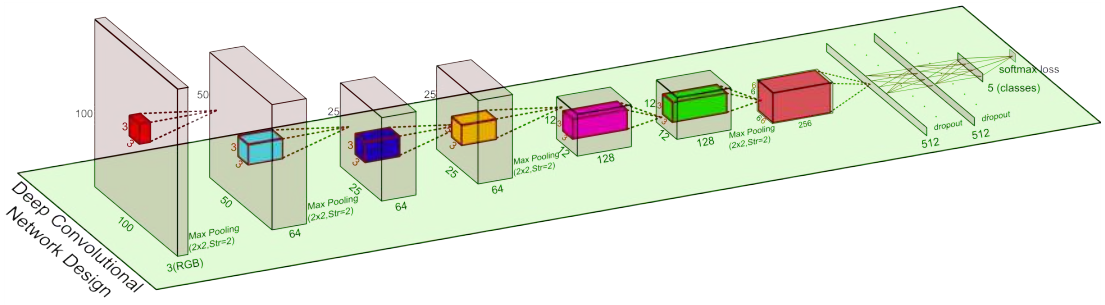


Figure 4.6: Designs of the Convolutional Neural Network. The design in Figure 4.3 is extended and this one is also motivated by [1]. All the convolution layers are followed by a ReLU normalization layer.

data set contains approximately ten times as many pixels as the training set of the detection part. But the network design lacks an improvement in order to discriminate multiple classes which necessitate capturing more complicated information. Therefore, the design in Figure 4.3 is upgraded with more layers, filters and neurons by sticking the style of [1] and regarding the base components.

The enhanced network is not a fully convolutional network and accept $100 \times 100 \times 3$ fixed sized inputs. Inputs are normalized by subtracting global mean of three channels. The network consists of 6 CONV layer with 3×3 filter size followed by 3 FC layer and concludes with softmaxloss layer. Except softmaxloss layer, all layers are followed by ReLU. CONV layers possess 64, 64, 64, 128, 128, 256 amount of filters in respective order and 1st, 2nd, 4th, and 6th layers are followed 2×2 max pooling operation with stride of 2. FC layers contain 512, 512 and 5 neurons (the number 5 is due to the number classes) and the first two layers are followed by dropout operation with 0.5 probability. Size of intermediate volumes between layers can be seen in the illustration of this design shown by Figure 4.6 and receptive fields of CONV layers are demonstrated in Figure 4.7.

4.2.3 Post-processing for Whole Slide Classification

The above network inputs a fixed size patch and produces class probabilities for it. In order to obtain probability maps for the whole slides, we need to

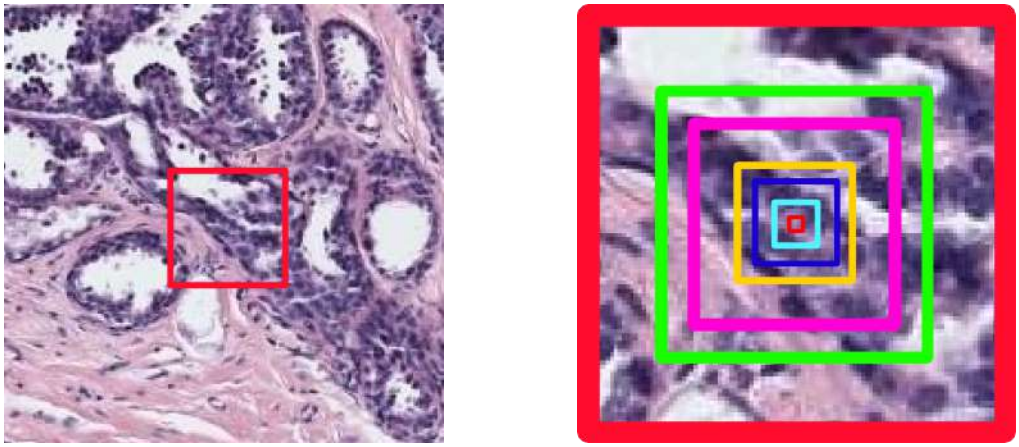


Figure 4.7: (left): The window size that the classification network is sensitive to, shown on a biopsy image. (right): Receptive fields of CONV layers shown in Figure 4.6 overlaid on 10X magnification, size of 100×100 pixels breast biopsy. Colors are selected according to their corresponding layer in Figure 4.6.

either classify patches extracted by sliding windows, or fully-convolutionalize the network. We do the latter as it allows more efficient WSI classification which, in fact, implicitly implements sliding windows with step size of 16. Therefore, each pixel of the probability maps corresponds to a 16×16 patch in the input space.

The probability maps produced by the network are first downsampled with bicubic interpolation in order to remove noise and regard only group of neighboring pixels with high probabilities in the original scale. Downsampled maps are then used to determine the final decision such that every pixel voted for the class that has the greatest probability for the pixel. Finally, the class with the majority of the votes is elected as the final prediction for the corresponding WSI.

Chapter 5

Experiments and Results

This chapter presents the experiments performed for both detection (Section 5.1) and classification tasks (Section 5.2). At the end of the chapter, we apply several visualization techniques in order to understand the visual information captured by our CNN network.

For both tasks, the data set is divided into 180 training and 60 testing images. After extracting the two set of training patches as explained in Chapter 4, we subdivide them into training and validation sets with 80% and 20% ratios respectively. Validation set is useful to observe when memorization of training set occurs, which is a common problem in deep learning. In our experiments, since we use dropout method, we did not face any overfitting issue.

CNN and FCN implementations are derived from the MatConvNet library [57] with a number of significant modifications and the networks are trained on a system that contains NVIDIA GeForce GTX 970 GPU, Intel(R) Xeon(R) E5-2630 v2 2.60GHz CPU and 64GB of RAM.

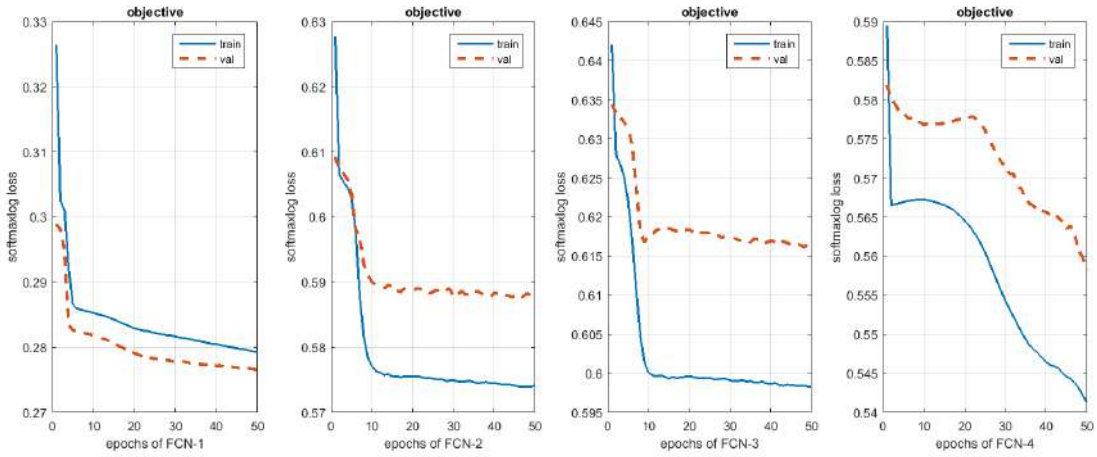


Figure 5.1: Learning curves of the four FCNs. Blue and red lines show softmaxlog loss of training and validation samples respectively. Usually, it is expected to obtain less error on the training set than the validation set, but FCN-1's curves show the otherwise. This may be due to trick used by MatConvNet library that is aggregating errors of the training set right after mini-batch updates for efficiency which should be done at the end of an epoch by calculating the error all-in-once.

5.1 ROI Detection

As the learning curves show in Figure 5.1, we train the four fully convolutional networks for 50 epochs with the training set extracted. For each FCN, the stochastic gradient descent algorithm is run to optimize total of 168,290 network parameters on mini batches of 25[1] images with 0.0001 learning rate, 0.0005 weight decay and 0.9 momentum. Those hyperparameters are empirically found on a subset of training data.

5.1.1 Reference Data

Since whole slide saliency detection on breast histopathology is not deeply investigated in the past, it is no surprise that there is no well-annotated data set for this task. Therefore, we develop a method to annotate the data set mentioned above for effective evaluation by using the given tracking information.

[1]We keep this number relatively low due to memory limitations

First, as in study [2], we use an assumption of saliency that consists of a set of rules to extract ground truth data set from screening logs of expert pathologists. A window is considered salient only if one of the following rules is satisfied:

- An expert zoomed into a region from the previous window and zoomed out right after.
- An expert waited in the same region for a signification amount of time (which is 2 seconds in the current settings).
- An expert slightly slide spatially on that region.

The algorithm until this point is directly adopted from the study [2] and all the implementation details are kept the same, thus we refer the readers to the study for more information. For each image, we take the union of all windows that satisfy the rules and create the proper label mask.

Secondly, we remove empty (white) regions that touch the boundaries from the mask with several morphological operations. The original image is first converted to Lab color space and then, thresholded at the value of 241 of the L channel. The resulting binary image is processed with morphological opening, closing and opening operations in this order with 3, 4, 6 radius disks as structuring elements. Then, connected components of the image are extracted and the components that touch boundaries are excluded from the mask formed in the first step. As a result, the empty (white) regions that contain no information are disregarded during the evaluation.

5.1.2 Evaluation Criteria

For each testing image, we obtain resulting probability map which is between 0 and 1 from the proposed detection pipeline and generate a binary ground truth by the above methodology. Additionally, for comparison, we attain the resulting saliency probabilities of the method proposed by [2] for the same images.

In order to measure effectiveness of the proposed approach, we first threshold the two probability maps at various levels and then, resulting binary masks are compared with the ground truth by counting pixels to calculate the confusion matrices for each threshold value. We draw some statistical curves using those confusion matrices such as Precision - Recall and Receiver operating characteristic (ROC) by averaging those values for 60 test images.

Theoretical efficiency estimations of this gradual pipeline can be viewed in Table 5.1 where we show the correlation of computational cost at different FCNs for the values of threshold ratio $\tau = 0, 0.1, 0.2 \ldots 0.9$. Let us say computational cost of processing a unit image in FCN-1 (in 0.625X magnification) is 1. Relative costs are calculated analytically based on their resolution and the remaining thresholded area, i.e., the resolution is doubled in each step (corresponding to quadrupled image size) and τ percent of remaining region is disregarded in every step. Finally, we calculate the efficiency gained compared to processing an image in all levels without thresholding or in only the last FCN (which is in 5X magnification). For example, for $\tau = 0.6$, we would detect saliency of an image by progressive elimination around 20 times faster than we would do by processing the full image in $5X$ magnification. In practice, the efficiency gain may be less than this optimistic estimation because the regions above threshold may be distributed all over the slide. The ideal case corresponds to having a single square region that is left after thresholding but in practice, multiple connected components with arbitrary shapes may need to be processed by giving their bounding boxes as input to the subsequent FCN.

An alternative to this adaptive thresholding is having fixed threshold values on the detection scores. We did not prefer this approach since the values are difficult to tune where the optimal values might be varying dependent on particular cases. On the other hand, the current thresholding method is logical in a way of selecting regions that are relatively more salient than others. For example, on a white paper with a small text, the most salient region is the text. But when we add a red disk on the paper, suddenly, the text becomes non-salient compared to the disk. In a similar manner, we determine salient regions in whole-slides relatively by this adaptive thresholding.

Table 5.1: Efficiency table for varying τ parameter values. Computational cost in all levels of the proposed pipeline is shown regarding the thresholding τ percentage of the image in each step. We compare total computational cost of the pipeline (6th row) with cost needed without thresholding (7th row) and only in 5X magnification.

Threshold Ratios (τ)	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
Comput. by FCN-1	1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
Comput. by FCN-2	4	3.24	2.56	1.96	1.44	1	0.64	0.36	0.16	0.04
Comput. by FCN-3	16	11.664	8.192	5.488	3.456	2	1.024	0.432	0.128	0.016
Comput. by FCN-4	64	41.9904	26.2144	15.3664	8.2944	4	1.6384	0.5184	0.1024	0.0064
Total computation	85	57.7944	37.7664	23.5144	13.7904	7.5	3.7024	1.6104	0.5904	0.1624
Efficiency comp. to all (85)	1	0.6799	0.4443	0.2766	0.1622	0.0882	0.0436	0.0190	0.0069	0.0019
Efficiency comp. to only 5X (64)	1.3281	030	0.5901	0.3674	0.2155	0.1172	0.0579	0.0252	0.0092	0.0025

5.1.3 Results & Discussion

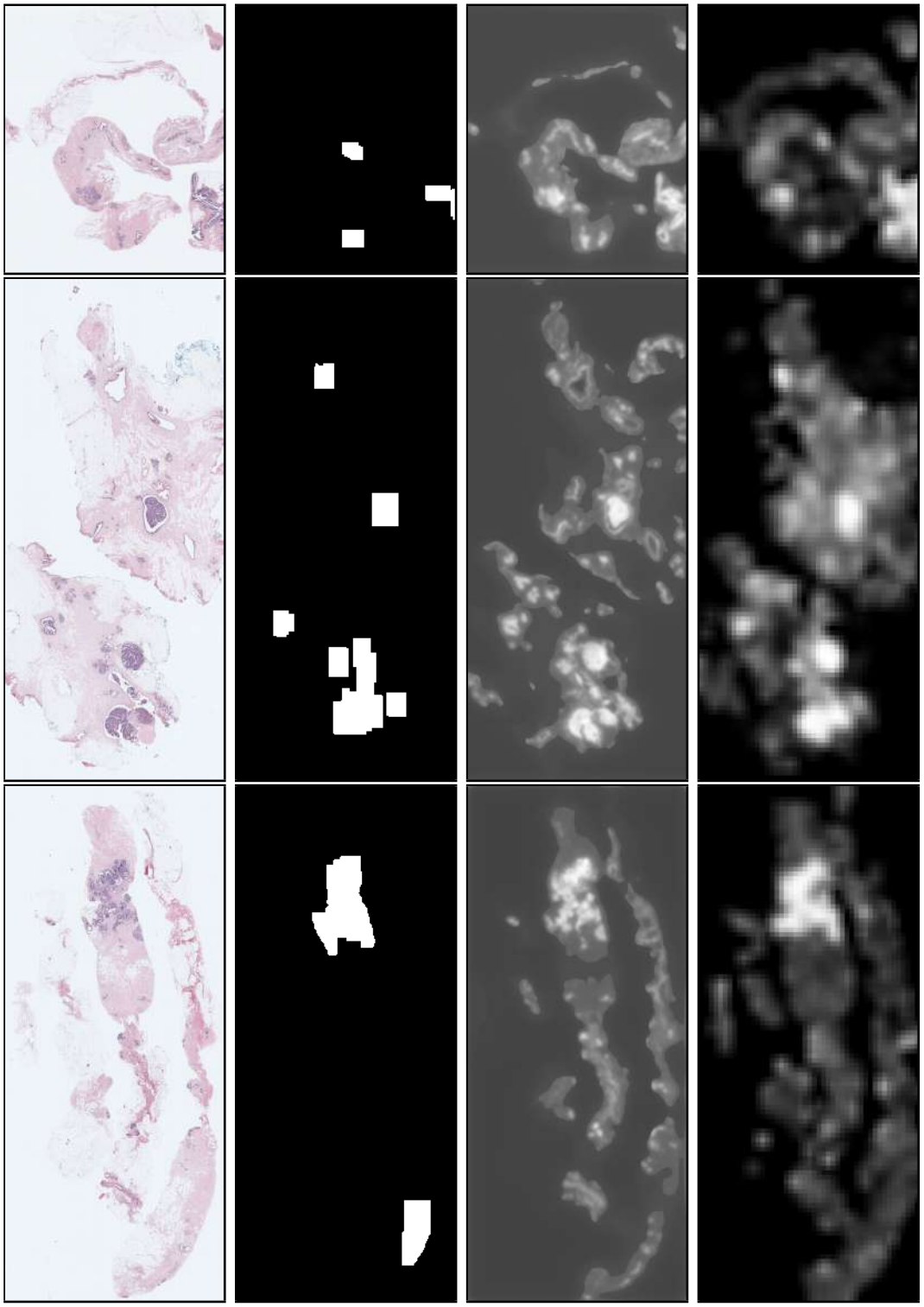
Figures 5.2, 5.3 and 5.4 show example images and the produced saliency maps. Figures 5.2 and 5.4 consist of the original images, corresponding ground truths, our saliency results and [2]'s saliency results. One may notice that the ground truth is not drawn at optimal precision, which might cause misleading results. We also see the extent of details achieved by the proposed method in Figure 5.4 whereas [2] produces more blurry results since it processes the image with sliding windows.

Thanks to the gradual elimination pipeline, while FCNs on higher resolutions capture the fine-details, they reduce the computation significantly on lower resolutions at the beginning. Moreover, in order to survive until the end of the pipeline, structures must consistently attain high probability in all levels because of the elimination of regions under a certain percent and taking the weighted geometric mean of outputs in all levels.

We show the potential of the proposed approach with Precision-Recall and ROC curves for different τ values in Figure 5.5 and 5.6, respectively. Since only a small portion of a WSI is considered as salient and we are interested to reduce the area to be processed in the following steps, it is more sensible to do our evaluation based on the false positive rate (FPR) rather than precision. For example, while doubling the precision may mean reducing the computation slightly, halving the

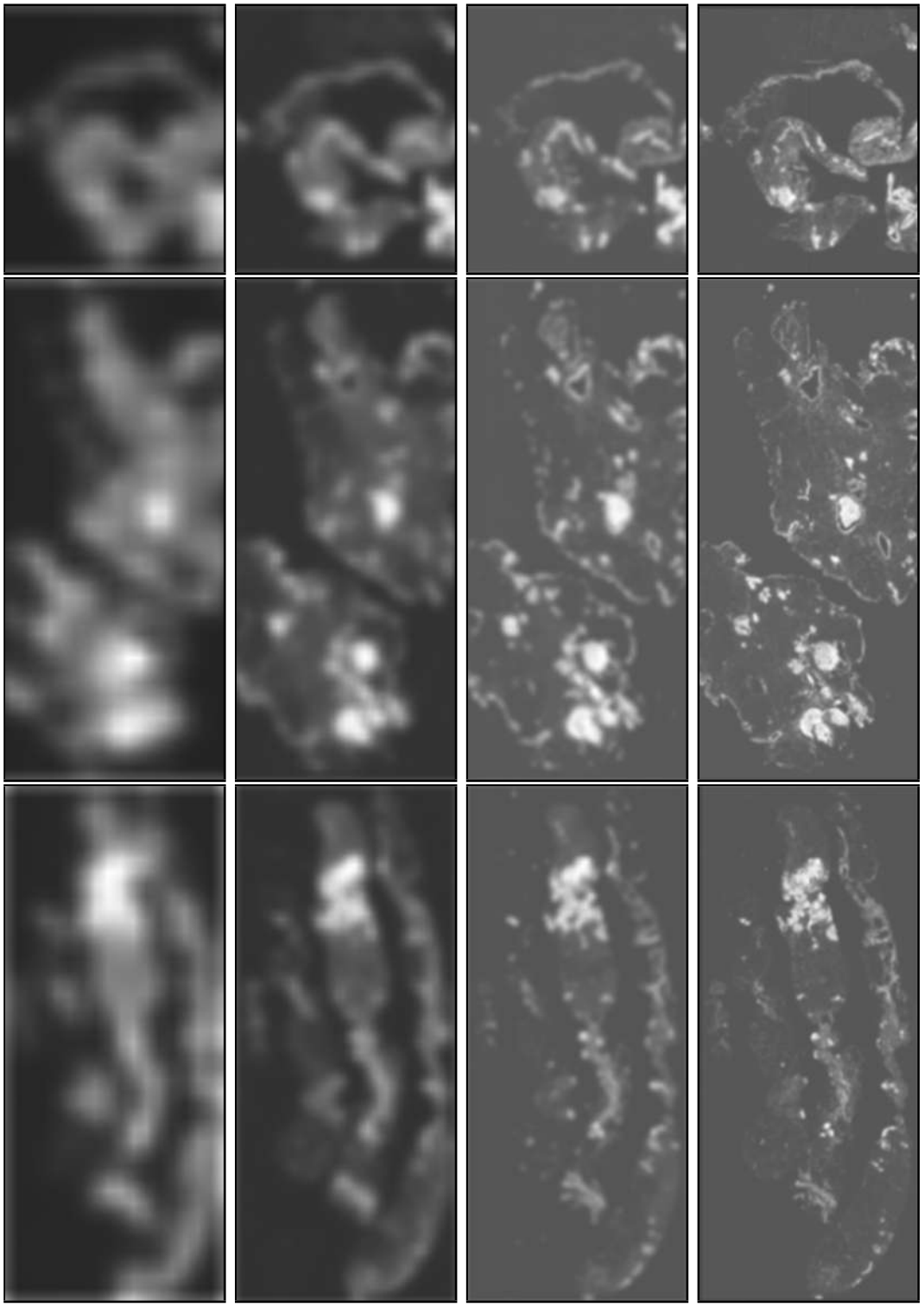
FPR means almost halving the computation as most areas of WSIs are non-salient. According to the Figure 5.6, while we observe improvement on both effectiveness and efficiency monotonically until $\tau = 0.4$, after that increasing τ values worsens the accuracy. So, there is an application dependent trade-off here, as higher τ values yield more efficiency according to the Table 5.1.

Our method attains the best area under curve (AUC) estimations of the ROC curves when $\tau = 0.3$ and $\tau = 0.4$ that are 0.9156 and 0.9153 respectively whereas [2] obtains an AUC score of 0.9124. We also see when $FPR = 0.2$, true positive rate (TPR) of [2] is 0.8552, while our method achieves TPR of 0.8947. Similarly, in the high TPR region above 0.8, our method results in smaller FPR values compared to [2]. Only after TPR of 0.98, [2] achieves higher TPR at the same FPR rate. Clearly, our method has slightly better performance than [2], yet, one needs to understand two handicaps our method is facing against [2]: (1) The proposed approach processes only a small portion of the image at utmost 5X magnification with the efficient pipeline where study [2] uses the whole images with 40X magnification versions which causes higher computational costs. (2) Note that the ground truth is generated by using the same saliency assumption made by [2]. Although, the assumption is just the beginning of their proposal, it should give remarkable advantage. Thus, despite the mentioned disadvantages, beating [2] shows the proposed saliency detection approach is promising.



[(a) (b) (c) (d)]

Figure 5.2: Example WSIs for saliency detection. (a) The original images. (b) The generated ground truth masks. (c) Resulting saliency maps of the proposed approach (Θ) for $\tau = 0.4$. (d) Outputs of the study [2]. *Best viewed with zoom.*



[(a) (b) (c) (d)]

Figure 5.3: The resulting saliency maps for $\tau = 0$ which are produced by each of the FCNs corresponding to the same images used in Figure 5.2. (a) Θ_1. (b) Θ_2. (c) Θ_3. (d) Θ_4. *Best viewed with zoom.*

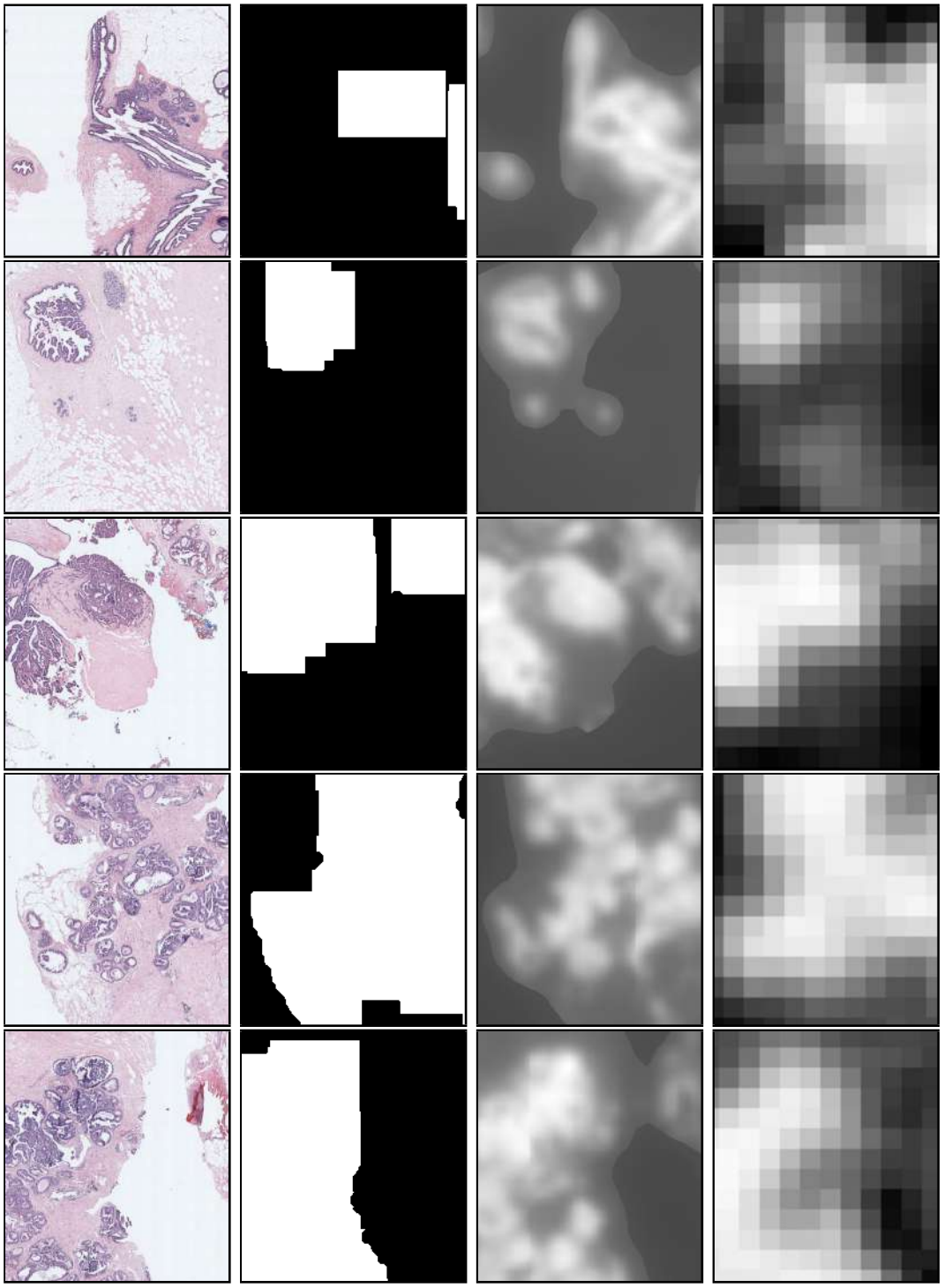


Figure 5.4: Cropped and zoomed samples extracted from the WSIs shown in Figure 5.2. (a) The original images. (b) The generated ground truth masks. (c) Resulting saliency maps of the proposed approach (Θ) for $\tau = 0.4$. (d) Outputs of the study [2]. *Best viewed with zoom.*

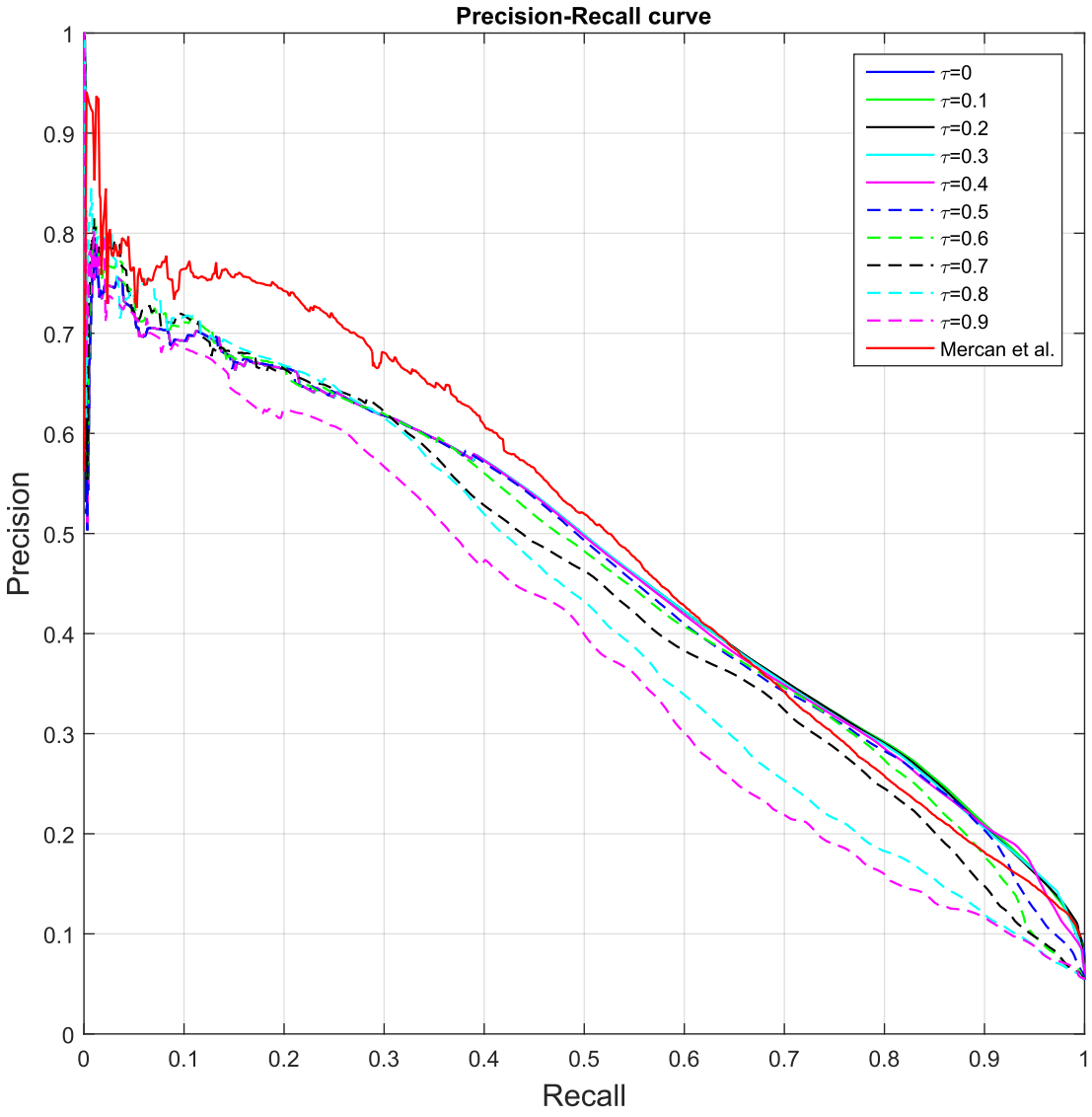


Figure 5.5: Precision-Recall curves of the proposed detection method with different τ values and result of the study [2].

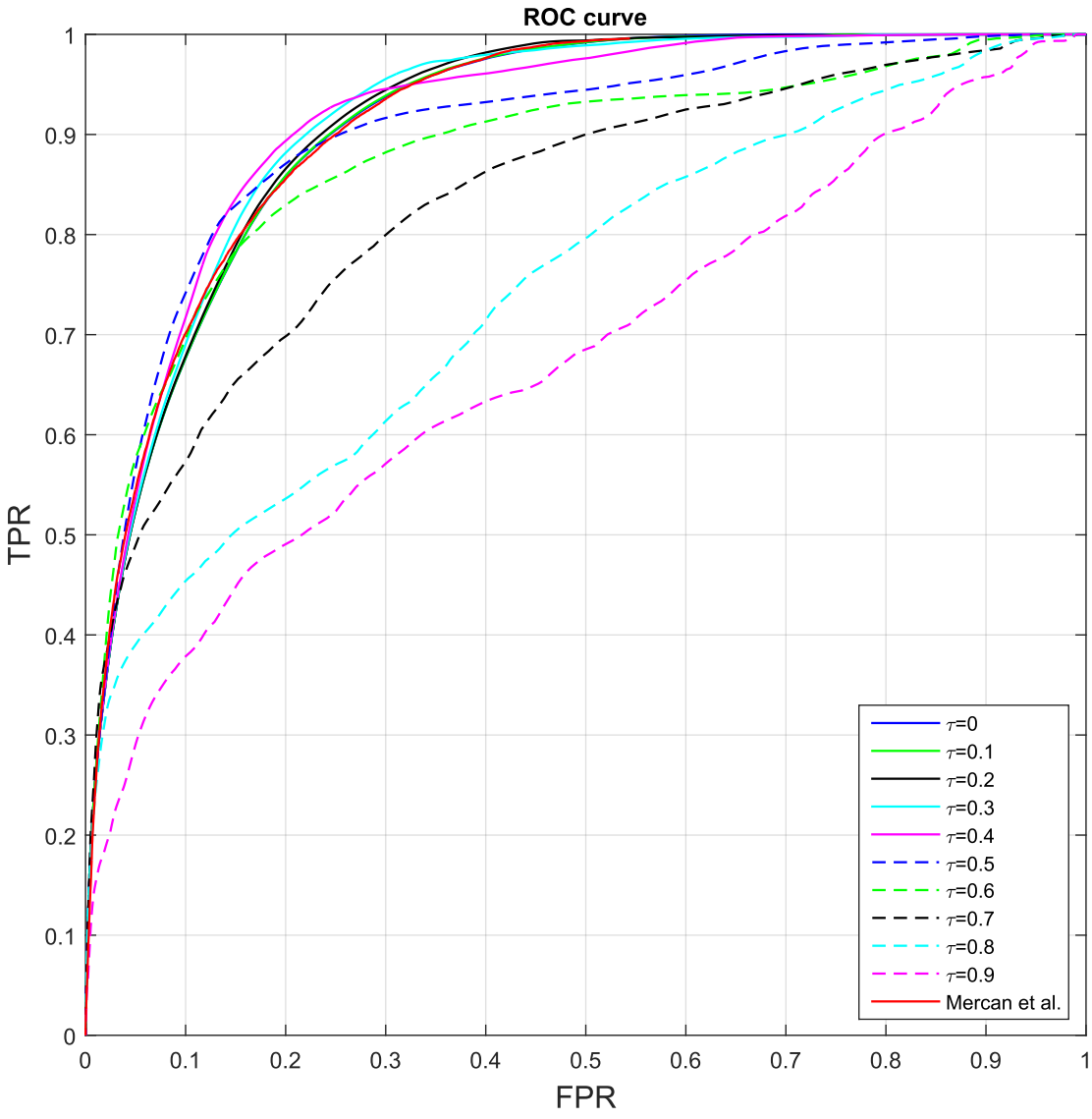


Figure 5.6: ROC curves of the proposed detection method with different τ values and result of the study [2].

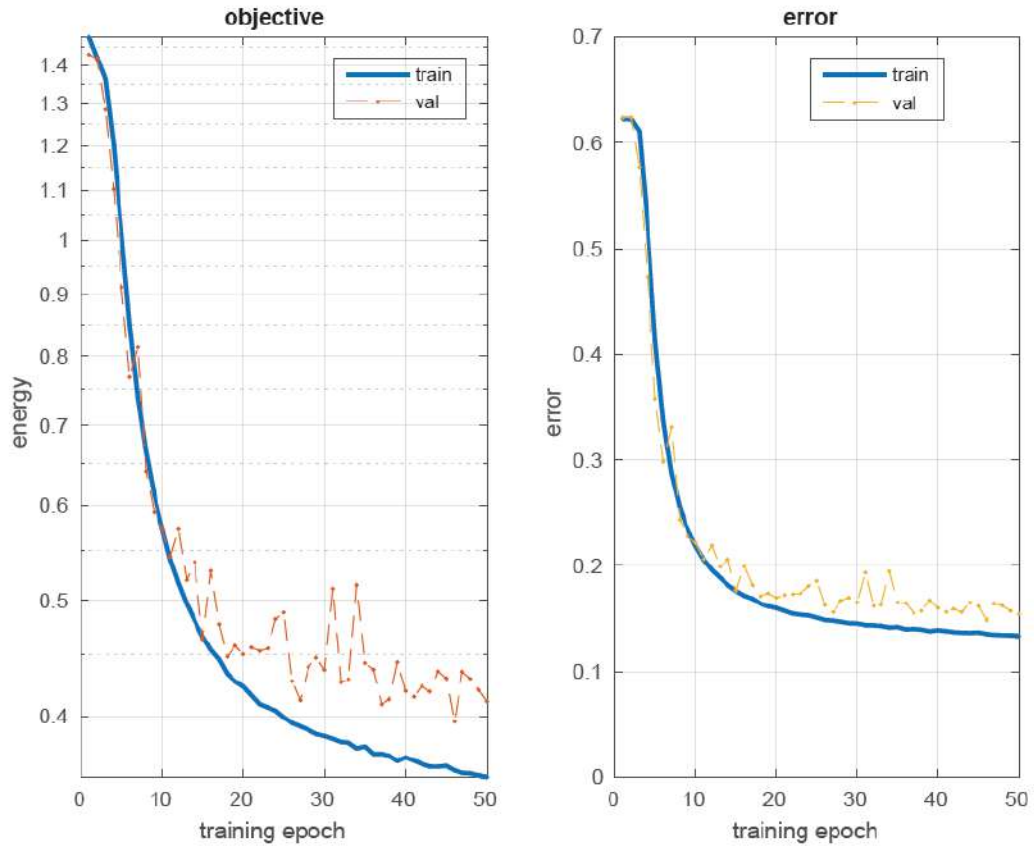


Figure 5.7: Learning curves of the CNN. Left figure shows the softmax loss of training and validation samples with blue and red lines, and the right figure shows the classification errors in a similar way.

5.2 ROI Classification

We train the CNN explained in Section 4.2.2 for the task of classification to five diagnostic cancer classes. As shown in Figure 5.7, the learning lasts 50 epochs to optimize total of 5,576,581 network parameters with mini batches of 256 images, 0.01 learning rate, 0.0005 weight decay and 0.9 momentum which are optimized on a small subset of the training data.

In order to evaluate the effectiveness of the trained CNN, we perform experiments for two tasks: classification of patches with size of 100×100 pixels and classification of WSI which requires the post-processing step.

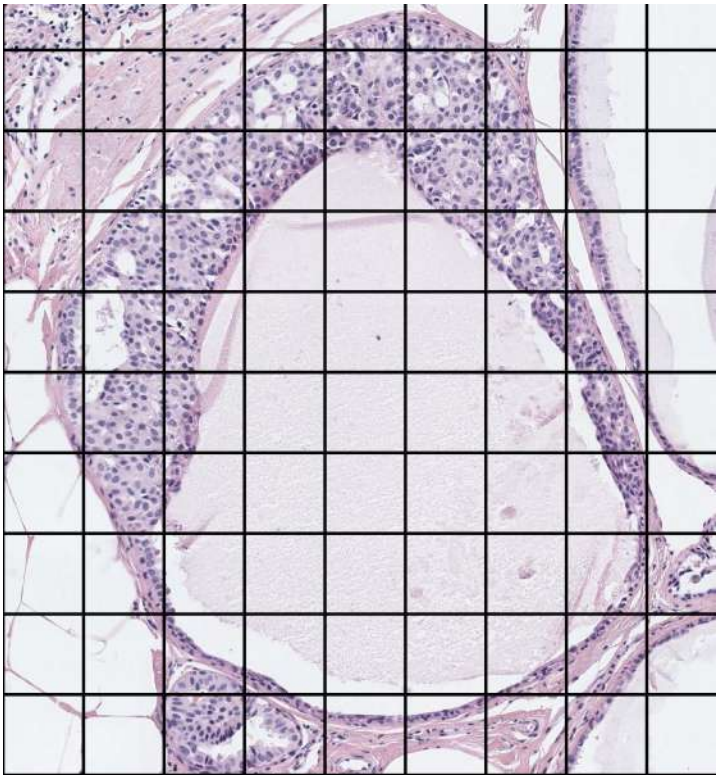


Figure 5.8: A region-of-interest size of 886 × 957 pixel marked on a WSI that is classified as ADH. The grid show the size of patches extracted.

5.2.1 Reference Data

The experiments to measure the performance of whole slide classification are performed on the same test set of 60 WSIs which is labeled by 45 pathologists and by 3 world-class expert pathologists completely. Consensus decision of the 3 pathologists is considered as the ground truth labels.

For the patch classification task, 209,654 image patches with size of 100 × 100 pixels are extracted from the 60 WSIs within the bounding boxes of the consensus annotations of the 3 expert pathologists. The distribution of the patches over classes is shown in Table 5.2. The patches are labeled with the ground truth label of the corresponding WSIs. Yet, the whole slides are labeled with the most critical diagnosis and the bounding boxes are drawn roughly, which means that patches that belong to less risky classes might be mislabeled. One such example is shown in Figure 5.8 where the majority patches belong to non-tissue regions.

Table 5.2: Distribution of the test data for patch classification task

Category	**Number of Samples**	**Percentage**
NP	10,401	4.96%
P	22,211	10.59%
ADH	34,151	16.29%
DCIS	92,107	43.93%
INV	50,784	24.22%
Total	209,654	100.00%

Table 5.3: Confusion matrix of the patch-based predictions of the proposed classifier for the test set.

		Predictions					*TPR*
	Classes	**NP**	**P**	**ADH**	**DCIS**	**INV**	*/Recall*
Ground Truth	**NP**	**2,477**	945	725	2,027	4,227	0.2381
	P	503	**7,246**	3,364	7,823	3,275	0.3262
	ADH	4,092	9,249	**5,727**	10,572	4,511	0.1676
	DCIS	5,003	23,074	7,068	**47,412**	9,550	0.5147
	INV	661	9,145	509	21,491	**18,978**	0.3737
FPR		0.0515	0.2263	0.0665	0.3566	0.1357	
Precision		0.1945	0.1459	0.3292	0.5307	0.4681	

5.2.2 Results & Discussion

5.2.2.1 Patch Classification

The accuracy of our CNN for classification of the extracted patches into 5 diagnostic classes is 39.04%. Given the accuracy in the validation set is around 85%, this accuracy is significantly low. This leads us to consider the possibility of memorization since validation and training sets are extracted from the same WSIs and therefore the network might have overfitted somehow. The confusion matrix of the classification results (shown in Table 5.3 where 63.21% of the wrong classifications belongs to lower triangle), also supports the hypothesis that the ground truth is biased such that all patches are labeled as the most dangerous class of the corresponding WSIs whereas our classifier tends to classify those patches into their accurate classes. Thus, our algorithm does not necessarily underestimate diagnostic categories.

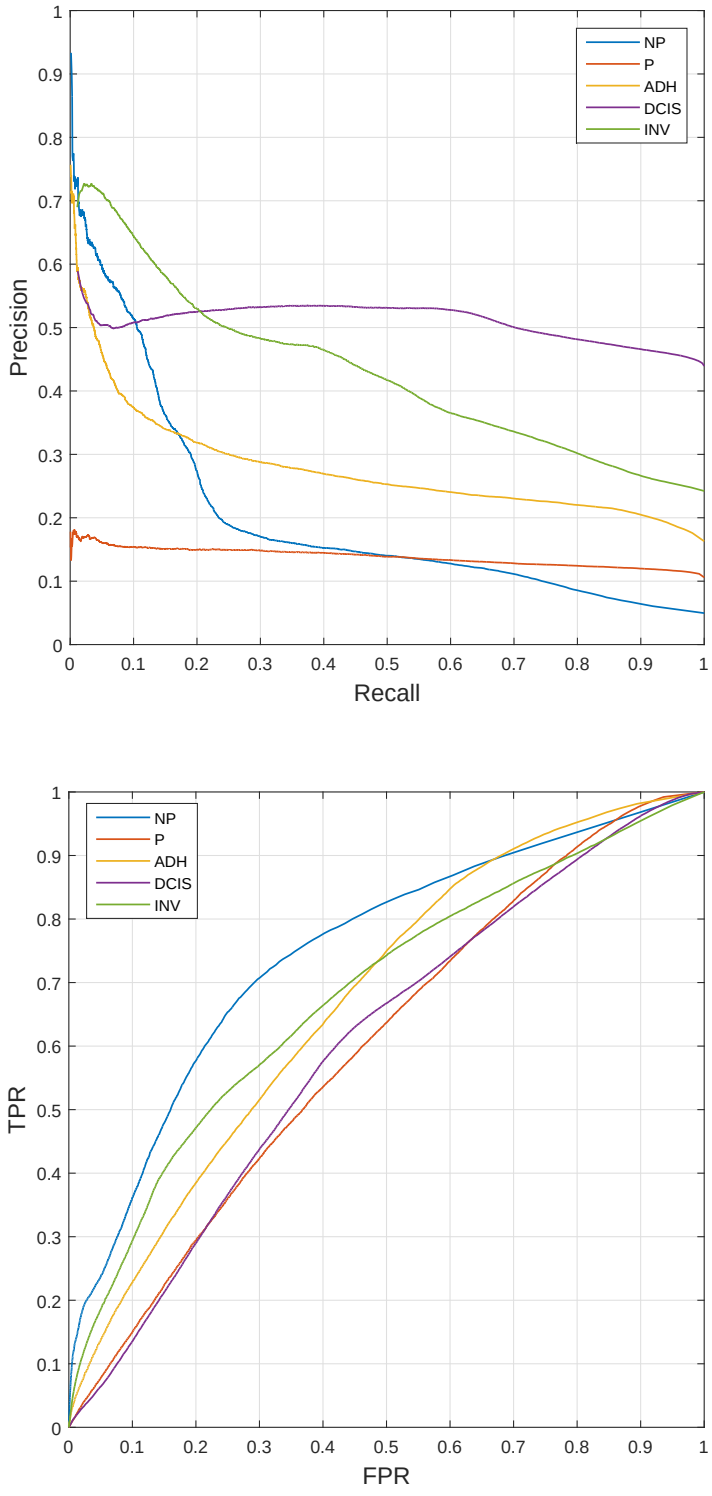


Figure 5.9: Precision - recall and ROC curves of the patch classification performance for five classes.

Other than predicted labels, we also have likelihood probabilities of each class for all the samples. For each of the five classes, we do binary classification individually for all the samples by thresholding their corresponding likelihood probabilities at different levels and draw five precision - recall and ROC curves as shown in Figure 5.9. This graphs supports the same argument mentioned above as we observe the DCIS and INV classes obtain 3-5 times higher precision than NP and P classes which are less critical for the patients and therefore, even if they exist in the ROI, more serious classes are preferred to them.

5.2.2.2 WSI Classification

For the classification of a WSI, we first obtain the class probability maps from the fully convolutionalized CNN which consist five probabilities for each pixel representing the class likelihoods and we downsample them by a factor of 7 as explained in Chapter 4. Second, as prior step, we extract the saliency map of the WSI as calculated in Section 5.1 and threshold it such that only the top 15% salient pixels remains. Then we find the classes that achieve the greatest probability at pixels detected as salient. The final class prediction of the WSI is decided by majority voting such that the most frequent winner class becomes the final prediction. The threshold parameter and the subsampling factor are optimized by maximizing the performance on the training set. Figures 5.10, 5.11, 5.12, 5.13, 5.14 show examples of slide-based classification for five WSIs including their winner class maps, saliency maps and probability maps for the five classes individually.

We evaluate the performance of the final predictions with the ground truth labels provided. Besides, we provide the performance when we exclude the saliency detection part in Table 5.4 where we also show the mean accuracy of the 3 expert and other 45 pathologists on both training and test sets. The confusion matrices and class-based TPR, FPR and precision scores of our CNN are shown in Tables 5.5 and 5.6 for training and test sets respectively.

The proposed system achieves a recognition rate of 55% which is comparable

Table 5.4: Classification accuracies of the slide-based predictions of the proposed method and the pathologists.

	Training Set	Test Set
3 Expert Pathologists	87.96 ± 3.57 %	87.78,± 2.55 %
Other 45 Pathologists	68.07 ± 10.36 %	65.44 ± 7.07 %
Proposed Approach	82.22 %	55 %
Proposed Approach w/o Saliency Detection	12.78 %	23.33 %

Table 5.5: Confusion matrix of the slide-based predictions of the proposed classifier for the training set.

		Predictions					*TPR*
	Classes	**NP**	**P**	**ADH**	**DCIS**	**INV**	*/Recall*
Ground Truth	**NP**	**5**	1	0	1	1	0.6250
	P	0	**45**	1	3	1	0.9000
	ADH	0	10	**32**	8	0	0.6400
	DCIS	0	3	0	**49**	3	0.8909
	INV	0	0	0	0	**17**	1
FPR		0	0.1051	0.0075	0.0936	0.0301	
Precision		1	0.7627	0.9697	0.8033	0.7727	

to the gold standard recorded by 45 human observers. A straightforward method for comparison just with the accuracies of individual pathologists shows us that our approach beats 4 out of 45 pathologists. It also performs better than or equal to 26 pathologists for ADH class, 21 for P class, 16 for NP class, 3 for DCIS class and no one for INV class.

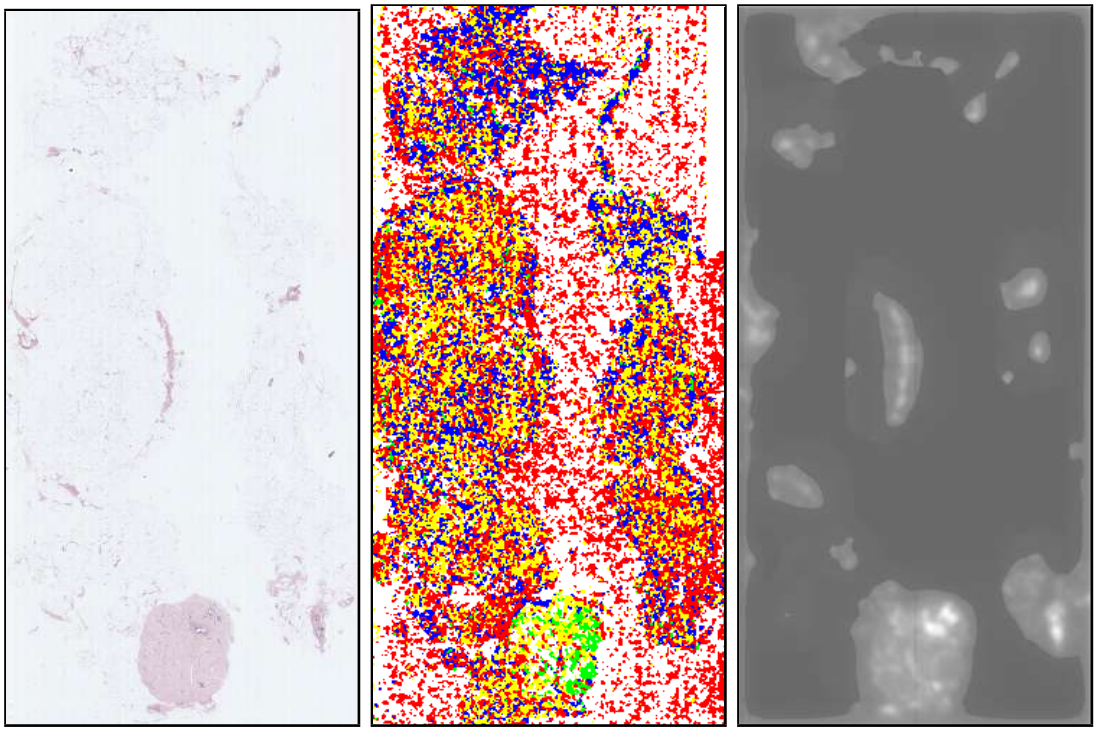
Additionally, we have applied McNemar's statistical test [58] to compare the proposed classifier with the other 45 pathologists' performance. Given the prediction of our classifier and diagnoses of the pathologists, we performed 45 McNemar tests where the null hypothesis is that there is no difference between the classifier and the pathologist. The tests are carried out with 5% significance level, and could reject the null hypothesis for 13 pathologists which means the differences between the performances are significant between our classifier and each individual. For the remaining 32 pathologists, we could not reject the hypothesis showing that we cannot claim their performances are significantly different than ours.

Table 5.6: Confusion matrix of the slide-based predictions of the proposed classifier for the test set.

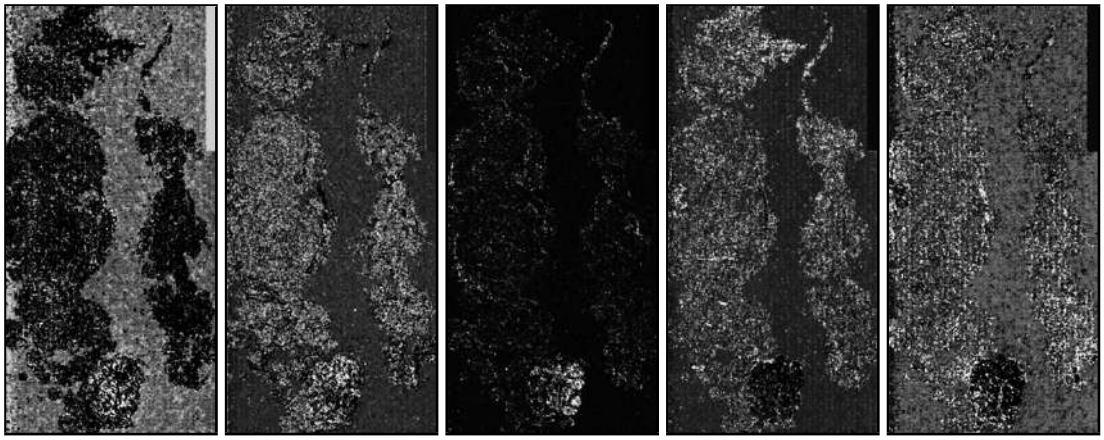
		Predictions					*TPR*
	Classes	**NP**	**P**	**ADH**	**DCIS**	**INV**	*/Recall*
Ground Truth	**NP**	**0**	2	0	1	2	0
	P	0	**9**	2	2	0	0.6923
	ADH	0	4	**4**	8	0	0.2500
	DCIS	0	2	0	**17**	2	0.8095
	INV	0	0	0	2	**3**	0.6000
FPR		-	0.1644	0.0434	0.3207	0.0705	
Precision		-	0.5294	0.6667	0.5667	0.4286	

Further, we performed a z-test with 10% significance where the null hypothesis is that our score belongs to the same normal distribution of the 45 pathologists' accuracies. We could not reject the null hypothesis which indicates our performance is somewhat similar to the pathologists.

These results indicate the hardness of the data set where the conflict among pathologists is also shown to be noticeably high by other studies and our method performs accurate enough to catch up the performance of human pathologists. Moreover, the effects of the prior saliency detection step that motivate its necessity are two-fold. First, we can avoid the computation of 85% of the image and improve efficiency around 6.6 times. Second, as shown in Table 5.4, the resulting performance is influenced significantly by means of effectiveness.

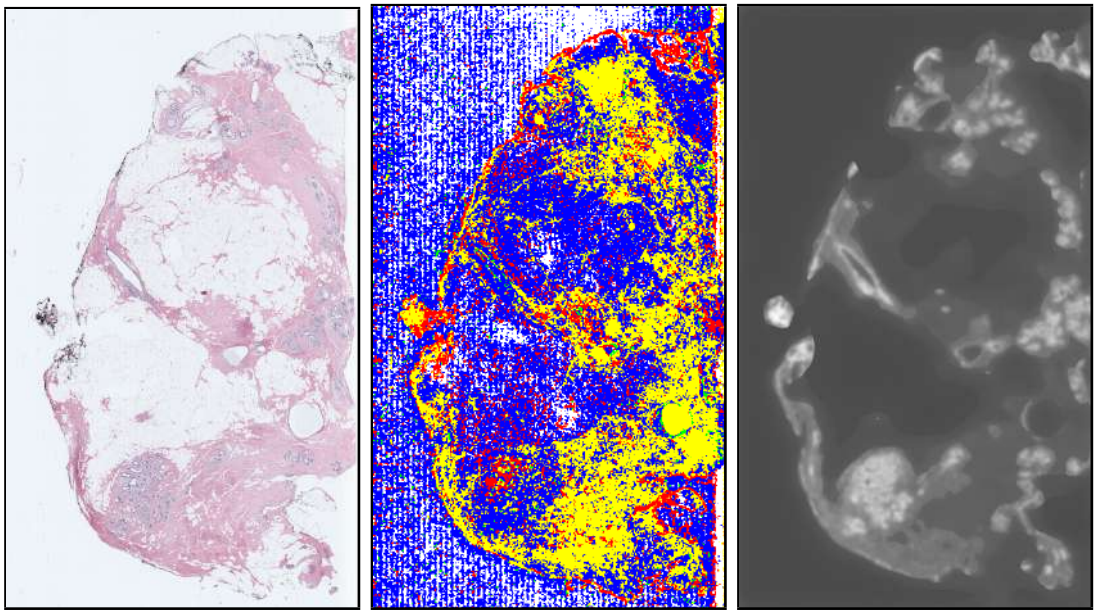


(a) The original Image (b) Winner Classes (c) Saliency Map

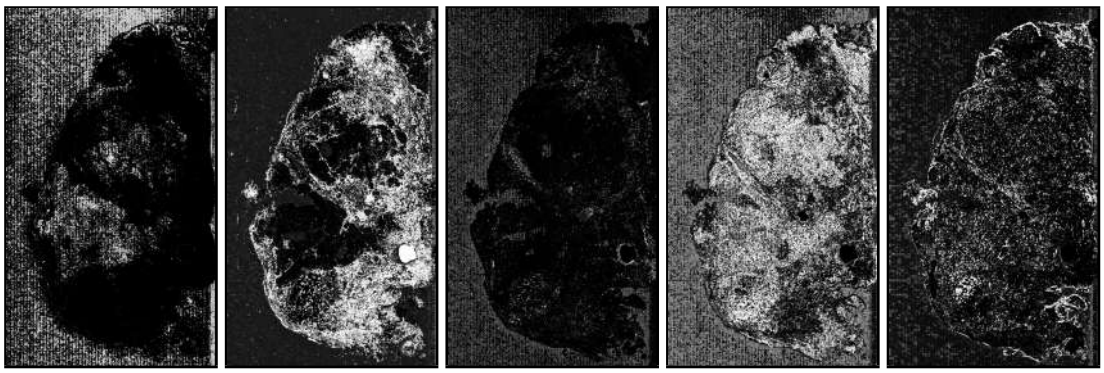


(d) Prob. of NP (e) Prob. of P (f) Prob. of ADH (g) Prob. of DCIS (h) Prob. of INV

Figure 5.10: A slide-based classification example of a WSI labeled as NP class. (a) The original Image. (b) Winner classes of each pixel where colors (i.e., white, yellow, green, blue, red) represent the five classes (i.e., NP, P, ADH, DCIS, INV respectively). (c) Corresponding saliency map obtained in Section 5.1. (d-h) Pixel-wise likelihood maps of five classes (i.e., NP, P, ADH, DCIS, INV respectively) according to the prediction of our classifier. This sample is wrongly classified as P after post-processing. *Best viewed in color and with zoom.*

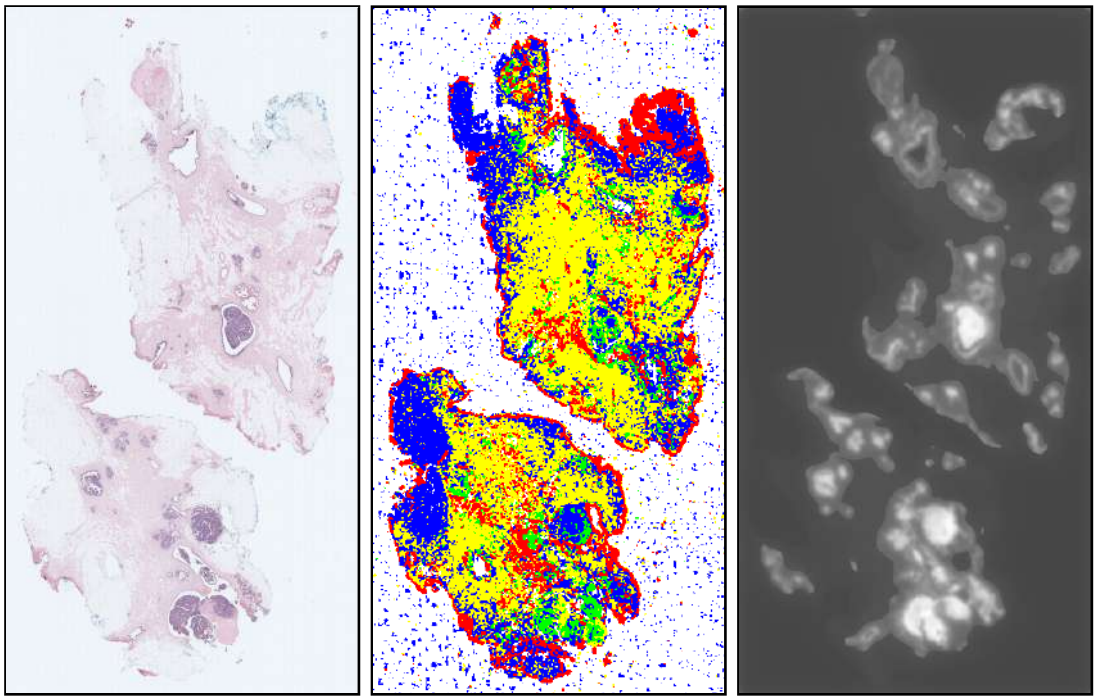


(a) The original Image (b) Winner Classes (c) Saliency Map

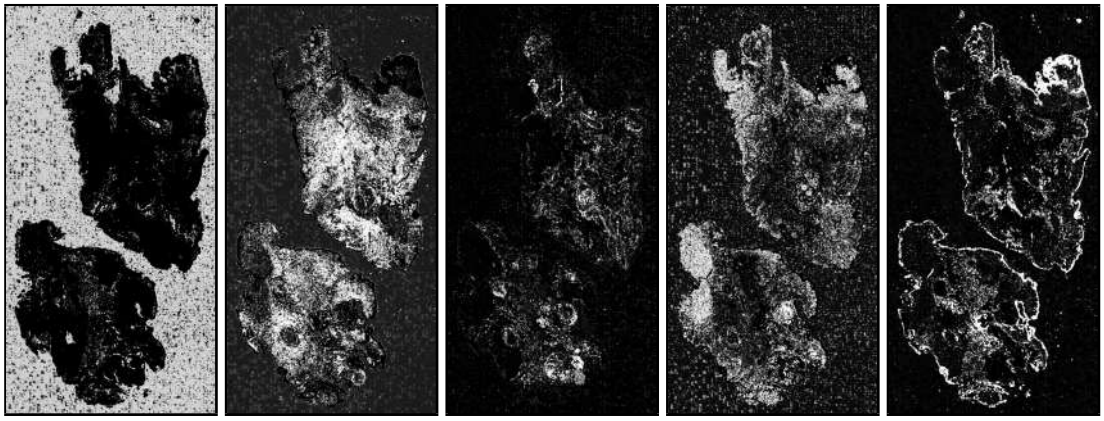


(d) Prob. of NP (e) Prob. of P (f) Prob. of ADH (g) Prob. of DCIS (h) Prob. of INV

Figure 5.11: A slide-based classification example of a WSI labeled as P class. (a) The original Image. (b) Winner classes of each pixel where colors (i.e., white, yellow, green, blue, red) represent the five classes (i.e., NP, P, ADH, DCIS, INV respectively). (c) Corresponding saliency map obtained in Section 5.1. (d-h) Pixel-wise likelihood maps of five classes (i.e., NP, P, ADH, DCIS, INV respectively) according to the prediction of our classifier. This sample is correctly classified as P after post-processing. *Best viewed in color and with zoom.*

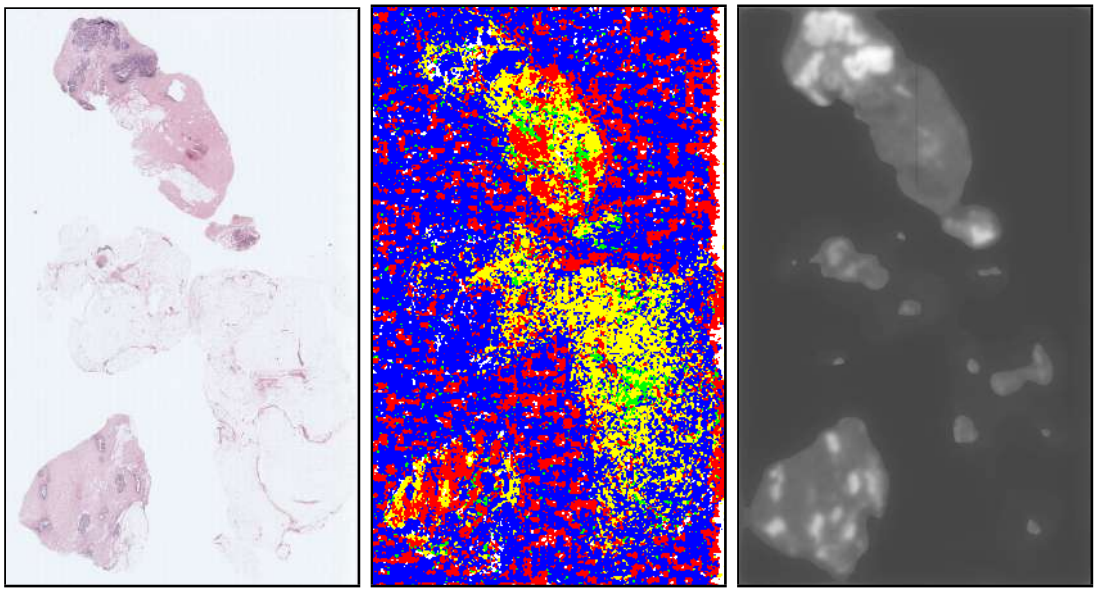


(a) The original Image (b) Winner Classes (c) Saliency Map



(d) Prob. of NP (e) Prob. of P (f) Prob. of ADH (g) Prob. of DCIS (h) Prob. of INV

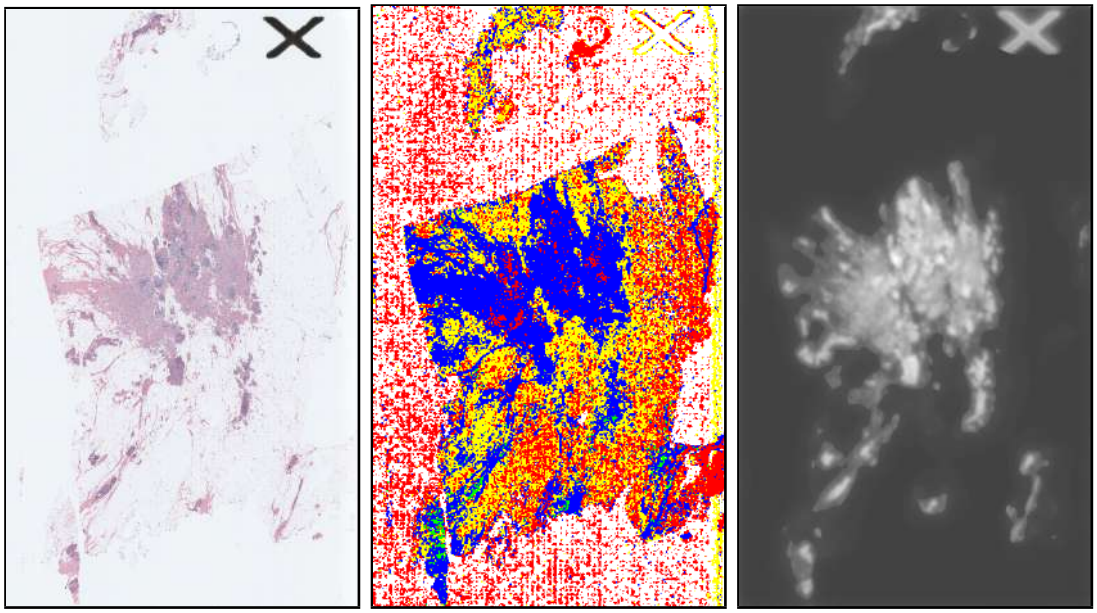
Figure 5.12: A slide-based classification example of a WSI labeled as ADH class. (a) The original Image. (b) Winner classes of each pixel where colors (i.e., white, yellow, green, blue, red) represent the five classes (i.e., NP, P, ADH, DCIS, INV respectively). (c) Corresponding saliency map obtained in Section 5.1. (d-h) Pixel-wise likelihood maps of five classes (i.e., NP, P, ADH, DCIS, INV respectively) according to the prediction of our classifier. This sample is correctly classified as ADH after post-processing. *Best viewed in color and with zoom.*



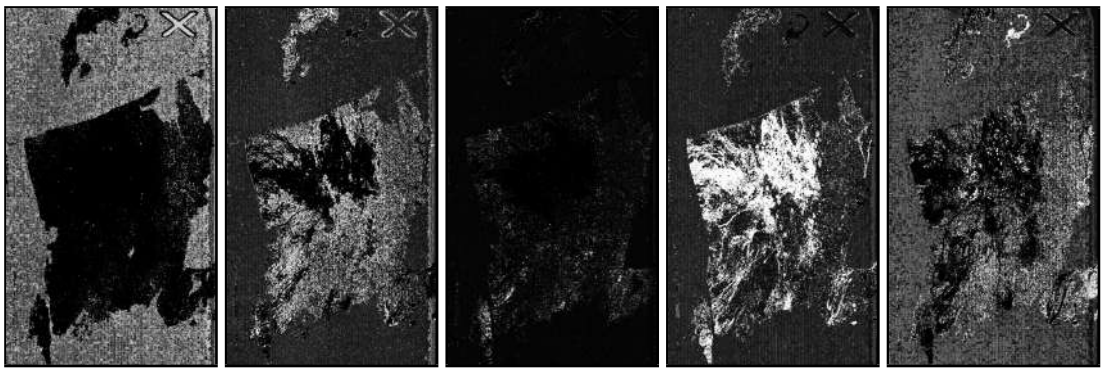
(a) The original Image (b) Winner Classes (c) Saliency Map

(d) Prob. of NP (e) Prob. of P (f) Prob. of ADH (g) Prob. of DCIS (h) Prob. of INV

Figure 5.13: A slide-based classification example of a WSI labeled as DCIS class. (a) The original Image. (b) Winner classes of each pixel where colors (i.e., white, yellow, green, blue, red) represent the five classes (i.e., NP, P, ADH, DCIS, INV respectively). (c) Corresponding saliency map obtained in Section 5.1. (d-h) Pixel-wise likelihood maps of five classes (i.e., NP, P, ADH, DCIS, INV respectively) according to the prediction of our classifier. This sample is correctly classified as DCIS after post-processing. *Best viewed in color and with zoom.*



(a) The original Image (b) Winner Classes (c) Saliency Map



(d) Prob. of NP (e) Prob. of P (f) Prob. of ADH (g) Prob. of DCIS (h) Prob. of INV

Figure 5.14: A slide-based classification example of a WSI labeled as INV class. (a) The original Image. (b) Winner classes of each pixel where colors (i.e., white, yellow, green, blue, red) represent the five classes (i.e., NP, P, ADH, DCIS, INV respectively). (c) Corresponding saliency map obtained in Section 5.1. (d-h) Pixel-wise likelihood maps of five classes (i.e., NP, P, ADH, DCIS, INV respectively) according to the prediction of our classifier. This sample is wrongly classified as DCIS after post-processing. *Best viewed in color and with zoom.*

5.3 Visualization

In the visualization experiments, we use the FeatureVis library [47] that implements occlusion and deconvolution methods as explained in [48] and [50] respectively. For activation maximization, mNeuron library [53] is adopted to produce synthesized images with the settings of [52].

We begin with the occlusion method that exposes the spatial effectiveness of the input image to the prediction by covering it partially with noise. As shown in Fig. 5.15, this visualization reveals critical information about the tissue structures that caused classifying or not classifying to particular classes. For example, 2., 4. and 9. rows show that intertwined nuclei may lead to classification of Invasive cancer. Last two rows are clear examples of why we need a prior saliency detection step since they are obviously empty regions, yet confusing the CNN. 7. and 10. rows contain wrong classification examples where a possible scenario is that the images are classified correctly but since the ground truth labels are marked according to the most dangerous class in the WSI, it seems to be misclassified.

Second, we show the top-9 responsive patches for each neuron and visualization of the contributions of their pixels in the input space in Figures 5.16, 5.17 and 5.18. The reconstructions are built by projecting the activations down to the input space with deconvolutional layers. One may notice that how filters develop abstract features based on simple patterns over the layers and how they group the patches with similar characteristic while expanding invariance such as shape and rotation. We also see that the low layers implements the fundamental features such as horizontal, vertical edges, T-junctions or blobs. Toward higher layers, we can observe various filters that are sensitive to particular arrangements of nuclei, lobules and duct structures. We leave the comment to the experts for further investigation of how similar CNN's features to those they are looking for during their clinical analysis.

Finally, we show the synthesized images that produce the maximum activation at the corresponding neurons in Figures 5.19 and 5.20. This technique starts with

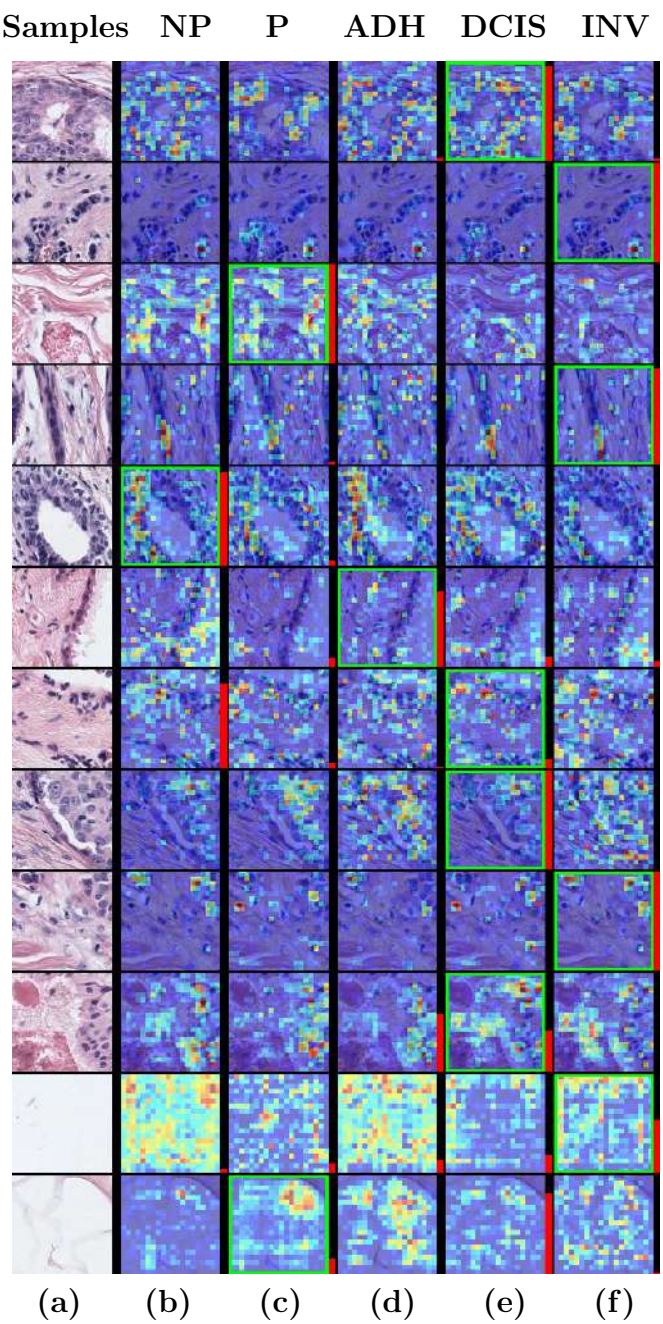


Figure 5.15: Classifications of several sample patches is visualized with occlusion method. (a) The original sample images with a size of 100 × 100 pixel. (b-f) Outputs of the occlusion method for five diagnostic classes overlaid on the original images. Ground truth diagnoses are indicated by green boxes. Predictions of our CNN are shown with red bar at the right side of the corresponding overlay. Warmer colors resemble higher effect of that region for the classification to the particular class. This effect may be either positive or negative, which means even if an image is not classified to a class, warmer colored pixels have the higher influence on that decision. *Best viewed in color and with zoom.*

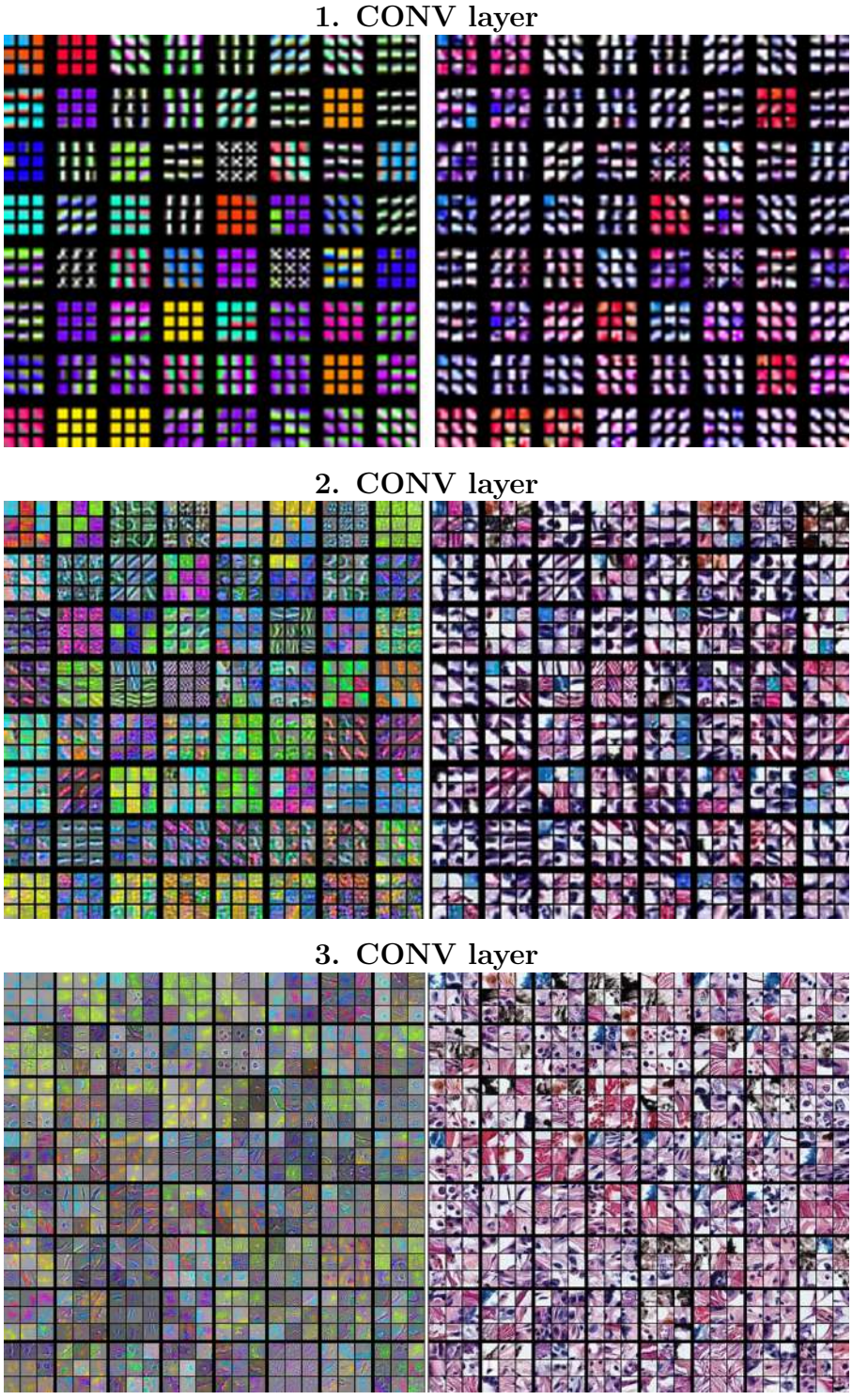
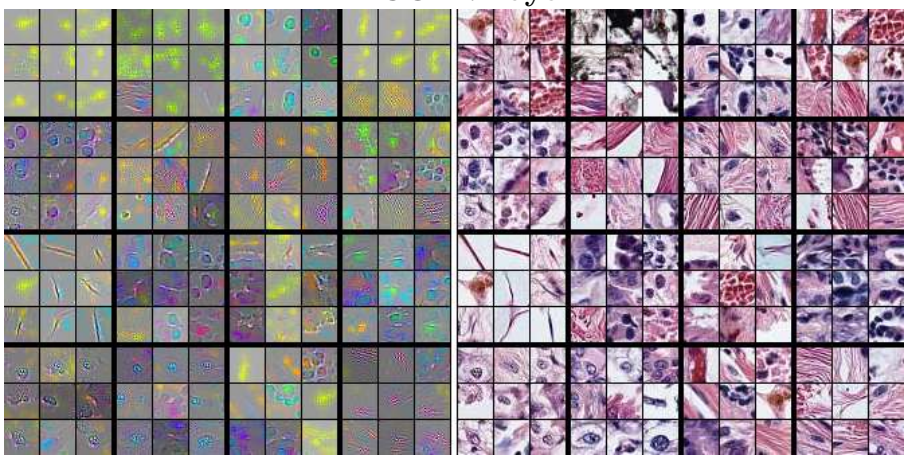
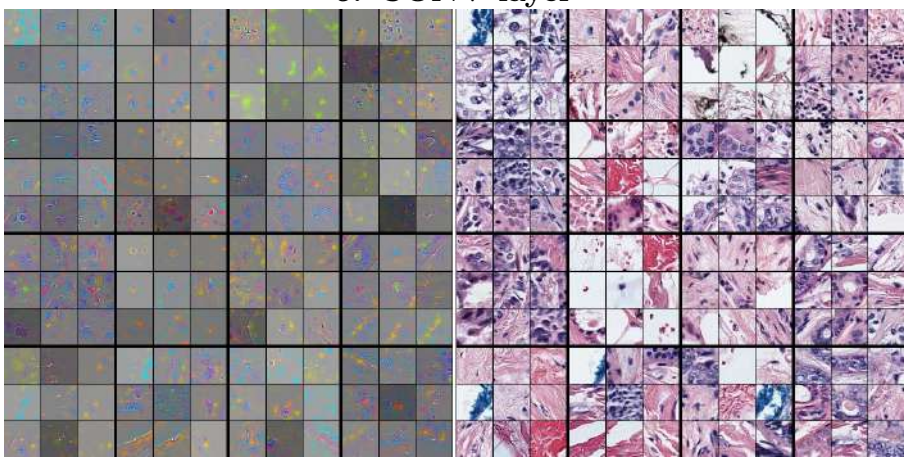


Figure 5.16: Visualization of the first three convolutional layers of the trained CNN model by the deconvolution method. For each layer, the top-9 activations of 64 neurons (left) and their corresponding original image patches (right) are shown as 3×3 groups. Color in left squares does not represent natural spectrum, but the activation projection and the color contrast artificially enhanced for better view. *Best viewed in color and with zoom.*

4. CONV layer



5. CONV layer



6. CONV layer

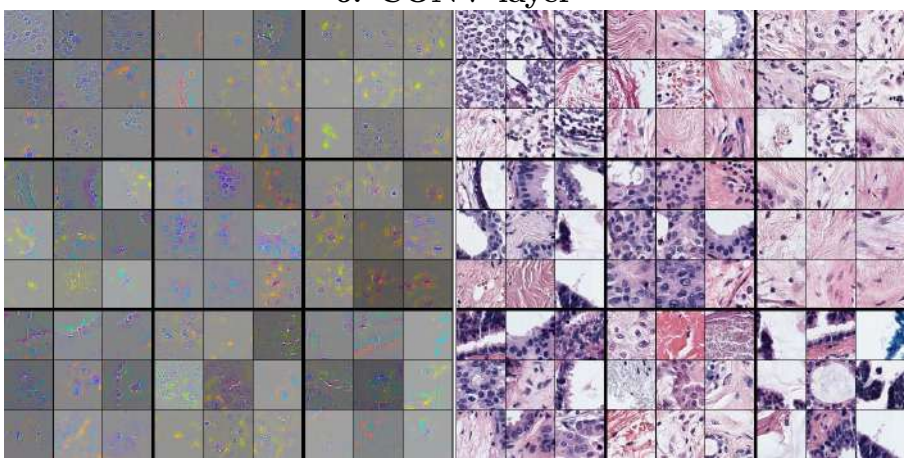
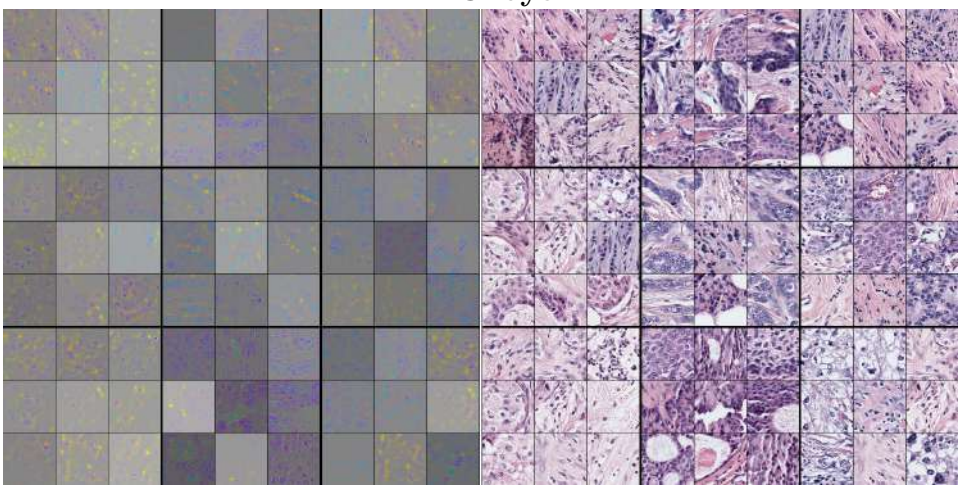
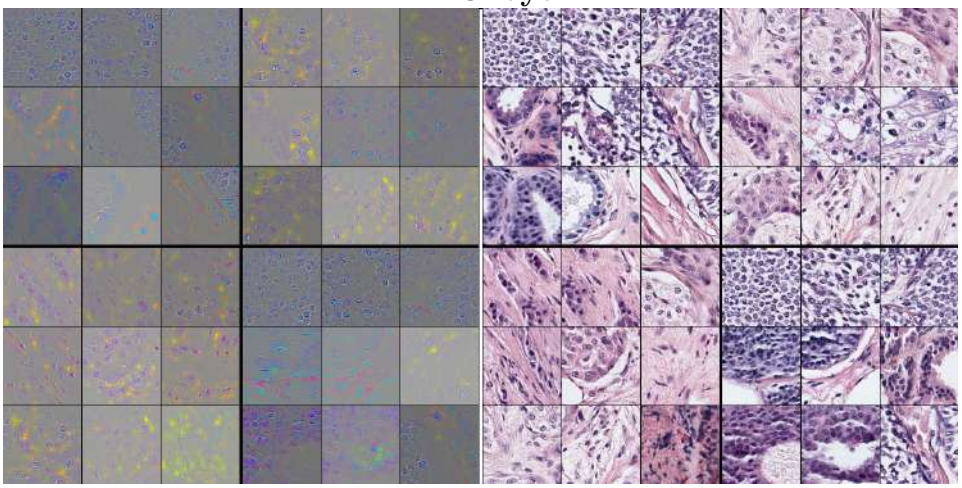


Figure 5.17: Visualization of the second three convolutional layers of the trained CNN model by the deconvolution method. For each layer, the top-9 activations of 64 neurons (left) and their corresponding original image patches (right) are shown as 3×3 groups. Color in left squares does not represent natural spectrum, but the activation projection and the color contrast artificially enhanced for better view. *Best viewed in color and with zoom.*

1. FC layer



2. FC layer



3. and the last FC layer

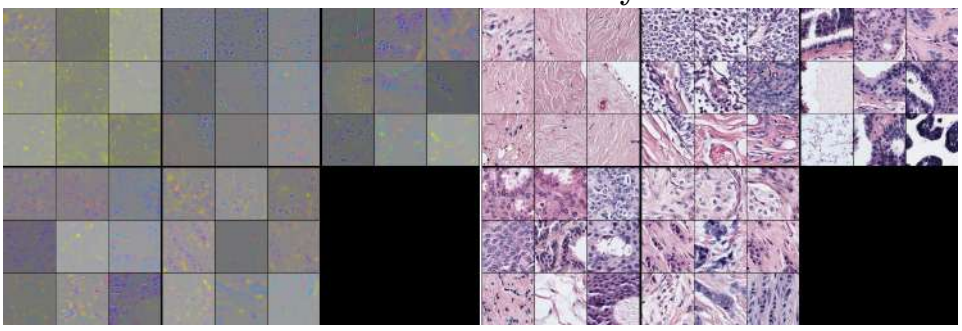
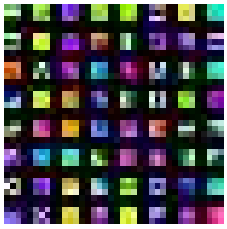
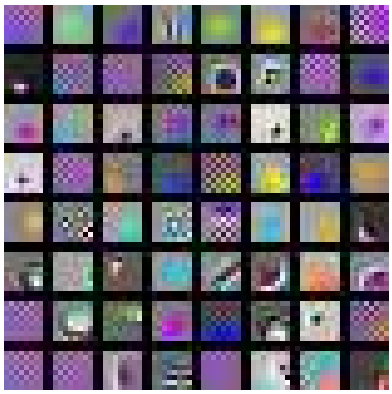


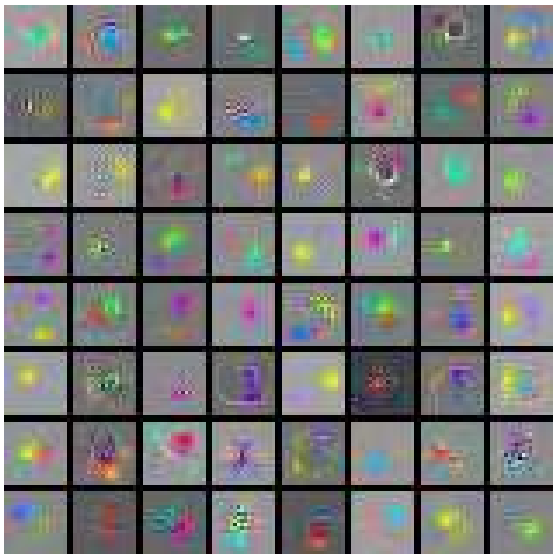
Figure 5.18: Visualization of the three fully-connected layers of the trained CNN model by the deconvolution method. For each layer, the top-9 activations of 64 neurons (left) and their corresponding original image patches (right) are shown as 3×3 groups. Color in left squares does not represent natural spectrum, but the activation projection and the color contrast artificially enhanced for better view. The last FC layer consists of five neurons that correspond to five classes. From top-left to bottom-right, the correspondence order is as follows: NP,P,ADH,DCIS,INV. *Best viewed in color and with zoom.*



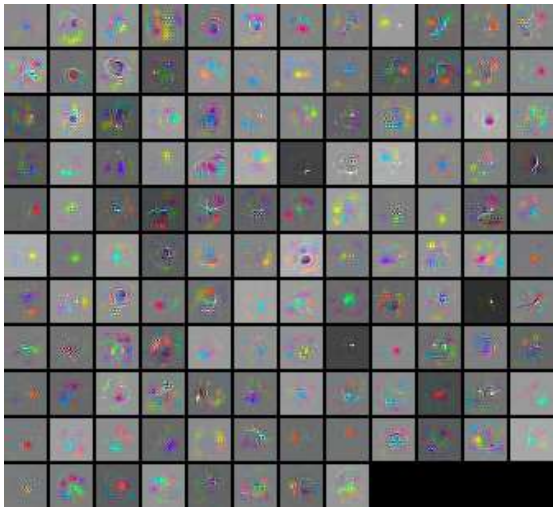
(a) 1. CONV layer



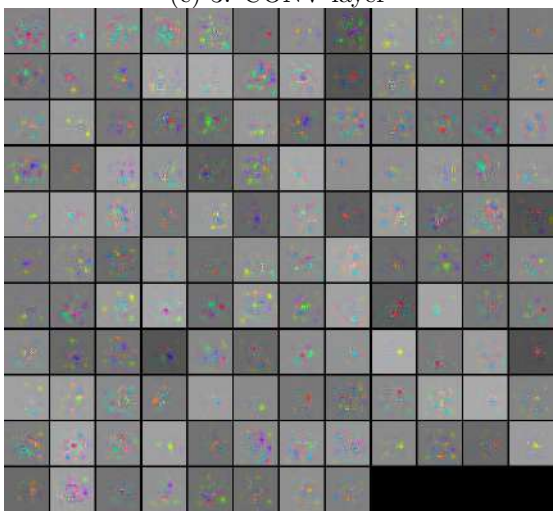
(b) 2. CONV layer



(c) 3. CONV layer

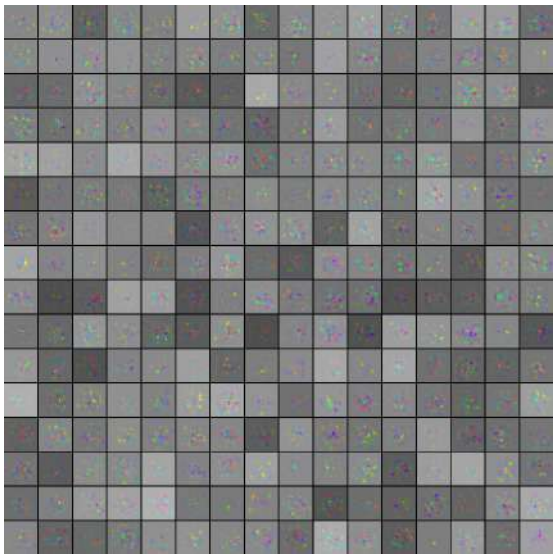


(d) 4. CONV layer

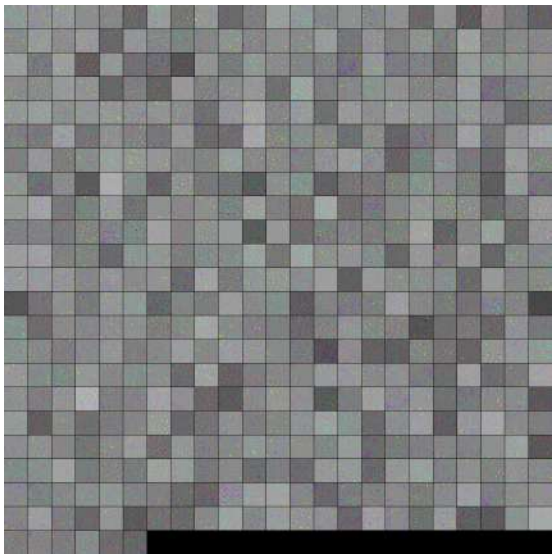


(e) 5. CONV layer

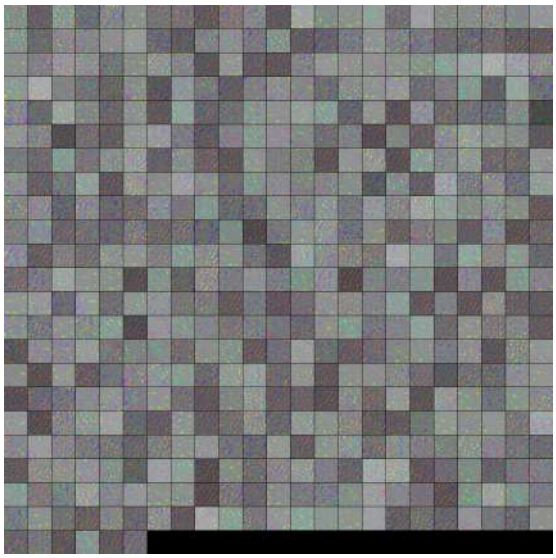
Figure 5.19: The synthesized images that activate neurons the most. *Best viewed in color and with zoom.*



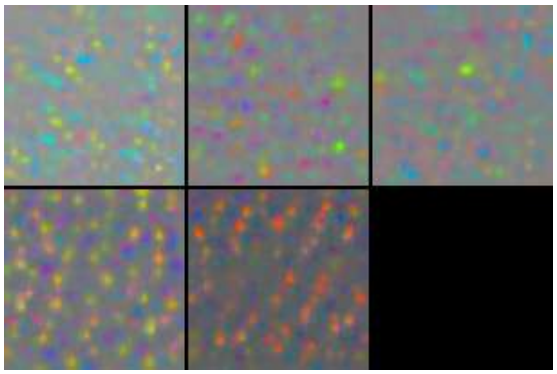
(a) 6. CONV layer



(b) 1. FC layer



(c) 2. FC layer



(d) 3. and last FC layer

Figure 5.20: The synthesized images that activate neurons the most. *Best viewed in color and with zoom.*

a white noise and updates it in the direction of negative gradient until convergence such that the activation of the neuron is maximized. This gives us intuition about the selectivity of the neurons. While we observe the varying patterns of neurons, the levels of abstraction increase toward the end of the network. We also noticed a number of dead neurons in the second layer which indicates less number of neurons could be sufficient in this layer.

Although, in theory, the activation maximization method should be the most straightforward method among the visualization methods, tuning its hyperparameters is its drawback since the quality of the resulting synthesized images are interpreted only by our visual objection which is not hundred percent reliable. Unfortunately, we fail to obtain natural images, yet the results are still giving insight about the filters' nature.

Chapter 6

Conclusion

In this thesis, we presented a complete CAD system for breast cancer diagnosis that inputs an RGB whole slide histopathology image and predict its cancer type over five diagnostic classes. The system includes: (1) a prior saliency detection step developed by training four sequential FCN models which imitates the way human pathologists perform diagnosis, (2) classification of the detected ROIs with the features learned by CNN, (3) a post WSI classification step that predicts a single diagnosis for one WSI out of the probability maps of classes. In other words, we inherently eliminate the healthy tissues in the detection part, and then classify the WSI according to the prediction of the majority class among the remaining cancerous regions.

We demonstrated the effectiveness and efficiency of the proposed approach on two comparisons: (1) obtained better saliency detection than the state-of-the-art method on the same data set, (2) approximation to the average accuracy of 45 human experts in WSI classification task where our method is not statistically different than the 32 experts with McNemar's tests.

Finally, we illustrate the visualizations of the features trained for the classification task that show that the features learned are purposeful and give intuition about what sort of features are discriminative.

In the future work, it could be studied how the proposed approach can be extended by including a more advanced post-processing for slide-based classification that consists of feature extraction and training another classifier for this task. We can also develop the networks further by simply enlarging them to enhance their receptive fields and learning capacity or by introducing multi-scale CNNs to capture textural and abstract information at the same time. Another direction would be adopting multi-instance multi-label learning since the whole-slides may include many regions at different diagnostic levels. Finally, a learned model may be transferred to other networks to initiate them instead of training all the networks from scratch.

Bibliography

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[2] E. Mercan, S. Aksoy, L. G. Shapiro, D. L. Weaver, T. Brunye, and J. G. Elmore, "Localization of diagnostically relevant regions of interest in whole slide images," in *22nd International Conference on Pattern Recognition (ICPR)*, pp. 1179–1184, IEEE, 2014.

[3] M. Veta, J. P. Pluim, P. J. van Diest, M. Viergever, *et al.*, "Breast cancer histopathology image analysis: A review," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 5, pp. 1400–1411, 2014.

[4] P. Boyle, B. Levin, *et al.*, *World Cancer Report.* IARC Press, International Agency for Research on Cancer, 2008.

[5] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *IEEE Reviews in Biomedical Engineering*, vol. 2, pp. 147–171, 2009.

[6] C. Mosquera-Lopez, S. Agaian, A. Velez-Hoyos, and I. Thompson, "Computer-aided prostate cancer diagnosis from digitized histopathology: A review on texture-based systems," *IEEE Reviews in Biomedical Engineering*, vol. 8, pp. 98–113, 2015.

[7] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[8] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, "Handwritten digit recognition with a back-propagation network," in *The Neural Information Processing Systems (NIPS)*, pp. 396–404, Morgan Kaufmann Publishers, 1990.

[9] Y. LeCun and M. Ranzato, "Deep learning tutorial," in *Tutorials in International Conference on Machine Learning (ICML)*, 2013.

[10] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[11] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 35, no. 8, pp. 1915–1929, 2013.

[12] C. Garcia and M. Delakis, "Convolutional face finder: A neural architecture for fast and robust face detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 26, no. 11, pp. 1408–1423, 2004.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, IEEE, 2009.

[14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[15] O. Matan, C. J. Burges, Y. LeCun, and J. S. Denker, "Multi-digit recognition using a space displacement neural network," in *The Neural Information Processing Systems (NIPS)*, pp. 488–495, Citeseer, 1991.

[16] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, 2015.

[17] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *International Conference on Learning Representations (ICLR)*, 2014.

[18] J. P. Bulte, L. Polman, M. Schlooz-Vries, A. Werner, R. Besselink, K. Sessink, R. Mus, S. Lardenoije, M. Imhof-Tas, J. Bulten, *et al.*, "One-day core needle biopsy in a breast clinic: 4 years experience," *Breast Cancer Research and Treatment*, vol. 137, no. 2, pp. 609–616, 2013.

[19] L. E. Boucheron, *Object-and spatial-level quantitative analysis of multispectral histopathology images for detection and characterization of cancer*. PhD thesis, University of California at Santa Barbara, 2008.

[20] S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features," in *5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 496–499, IEEE, 2008.

[21] A. Chekkoury, P. Khurd, J. Ni, C. Bahlmann, A. Kamen, A. Patel, L. Grady, M. Singh, M. Groher, N. Navab, *et al.*, "Automated malignancy detection in breast histopathological images," in *SPIE Medical Imaging*, pp. 831515–831515, International Society for Optics and Photonics, 2012.

[22] C. Gunduz, B. Yener, and S. H. Gultekin, "The cell graphs of cancer," *Bioinformatics*, vol. 20, no. suppl 1, pp. i145–i151, 2004.

[23] R. Albert, T. Schindewolf, I. Baumann, and H. Harms, "Three-dimensional image processing for morphometric analysis of epithelium sections," *Cytometry*, vol. 13, no. 7, pp. 759–765, 1992.

[24] S. Doyle, M. Hwang, K. Shah, A. Madabhushi, M. Feldman, and J. Tomaszeweski, "Automated grading of prostate cancer using architectural and textural image features," in *4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 1284–1287, IEEE, 2007.

[25] C. Bilgin, C. Demir, C. Nagi, and B. Yener, "Cell-graph mining for breast tissue modeling and classification," in *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 5311–5314, IEEE, 2007.

[26] C. C. Bilgin, P. Bullough, G. E. Plopper, and B. Yener, "Ecm-aware cell-graph mining for bone tissue modeling and classification," *Data Mining and Knowledge Discovery*, vol. 20, no. 3, pp. 416–438, 2010.

[27] E. Cosatto, M. Miller, H. P. Graf, and J. S. Meyer, "Grading nuclear pleomorphism on histological micrographs," in *19th International Conference on Pattern Recognition (ICPR)*, pp. 1–4, IEEE, 2008.

[28] S. Naik, S. Doyle, M. Feldman, J. Tomaszewski, and A. Madabhushi, "Gland segmentation and computerized gleason grading of prostate histology by integrating low-, high-level and domain specific information," in *MIAAB workshop*, pp. 1–8, Citeseer, 2007.

[29] C.-H. Huang, A. Veillard, L. Roux, N. Loménie, and D. Racoceanu, "Time-efficient sparse analysis of histopathological whole slide images," *Computerized Medical Imaging and Graphics*, vol. 35, no. 7, pp. 579–591, 2011.

[30] S. Doyle, A. Madabhushi, M. Feldman, and J. Tomaszeweski, "A boosting cascade for automated detection of prostate cancer from digitized histology," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 504–511, Springer, 2006.

[31] O. Sertel, J. Kong, H. Shimada, U. Catalyurek, J. H. Saltz, and M. N. Gurcan, "Computer-aided prognosis of neuroblastoma on whole-slide images: Classification of stromal development," *Pattern Recognition*, vol. 42, no. 6, pp. 1093–1103, 2009.

[32] J. Kong, O. Sertel, H. Shimada, K. L. Boyer, J. H. Saltz, and M. N. Gurcan, "Computer-aided evaluation of neuroblastoma on whole-slide histology images: Classifying grade of neuroblastic differentiation," *Pattern Recognition*, vol. 42, no. 6, pp. 1080–1092, 2009.

[33] A. Basavanhally, S. Ganesan, M. Feldman, N. Shih, C. Mies, J. Tomaszewski, and A. Madabhushi, "Multi-field-of-view framework for distinguishing tumor grade in er+ breast cancer from entire histopathology slides," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 8, pp. 2089–2099, 2013.

[34] B. E. Bejnordi, M. Balkenhol, G. Litjens, R. Holland, P. Bult, N. Karssemeijer, and J. A. van der Laak, "Automated detection of dcis in whole-slide h&e stained breast histopathology images," *IEEE Transactions on Medical Imaging*, vol. PP, no. 99, 2016.

[35] M. Peikari, M. J. Gangeh, J. Zubovits, G. Clarke, and A. L. Martel, "Triaging diagnostically relevant regions from pathology whole slides of breast cancer: A texture based approach," *IEEE transactions on medical imaging*, vol. 35, no. 1, pp. 307–315, 2016.

[36] C. Bahlmann, A. Patel, J. Johnson, J. Ni, A. Chekkoury, P. Khurd, A. Kamen, L. Grady, E. Krupinski, A. Graham, *et al.*, "Automated detection of diagnostically relevant regions in h&e stained digital pathology slides," in *SPIE Medical Imaging*, pp. 831504–831504, International Society for Optics and Photonics, 2012.

[37] A. Cruz-Roa, A. Basavanhally, F. González, H. Gilmore, M. Feldman, S. Ganesan, N. Shih, J. Tomaszewski, and A. Madabhushi, "Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks," in *SPIE Medical Imaging*, pp. 904103–904103, International Society for Optics and Photonics, 2014.

[38] Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, E. I. Chang, *et al.*, "Deep learning of feature representation with multiple instance learning for medical image analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1626–1630, IEEE, 2014.

[39] A. Cruz-Roa, J. Arevalo, A. Basavanhally, A. Madabhushi, and F. González, "A comparative evaluation of supervised and unsupervised representation learning approaches for anaplastic medulloblastoma differentiation," in *Tenth International Symposium on Medical Information Processing and*

Analysis, pp. 92870G–92870G, International Society for Optics and Photonics, 2015.

[40] Y. Song, L. Zhang, S. Chen, D. Ni, B. Lei, and T. Wang, "Accurate segmentation of cervical cytoplasm and nuclei based on multiscale convolutional network and graph partitioning," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 10, pp. 2421–2433, 2015.

[41] A. Cruz-Roa, J. Arévalo, A. Judkins, A. Madabhushi, and F. González, "A method for medulloblastoma tumor differentiation based on convolutional neural networks and transfer learning," in *11th International Symposium on Medical Information Processing and Analysis (SIPAIM 2015)*, pp. 968103–968103, International Society for Optics and Photonics, 2015.

[42] G. Litjens, C. I. Sánchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, C. Hulsbergen-van de Kaa, P. Bult, B. van Ginneken, and J. van der Laak, "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," *Scientific Reports*, vol. 6, 2016.

[43] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, "Deep learning for identifying metastatic breast cancer," *arXiv preprint arXiv:1606.05718*, 2016.

[44] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *IEEE Conference on European Conference on Computer Vision (ECCV)*, pp. 818–833, Springer, 2014.

[45] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks.," in *Aistats*, vol. 15, p. 275, 2011.

[46] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting.," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[47] F. Grün, C. Rupprecht, N. Navab, and F. Tombari, "A taxonomy and library for visualizing learned features in convolutional neural networks," *arXiv preprint arXiv:1606.07757*, 2016.

[48] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene cnns," *arXiv preprint arXiv:1412.6856*, 2014.

[49] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[50] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[51] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," Tech. Rep. 1341, University of Montreal, 2009.

[52] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5188–5196, IEEE, 2015.

[53] D. Wei, B. Zhou, A. Torrabla, and W. Freeman, "Understanding intra-class knowledge inside cnn," *arXiv preprint arXiv:1507.02379*, 2015.

[54] J. G. Elmore, G. M. Longton, P. A. Carney, B. M. Geller, T. Onega, A. N. Tosteson, H. D. Nelson, M. S. Pepe, K. H. Allison, S. J. Schnitt, *et al.*, "Diagnostic concordance among pathologists interpreting breast biopsy specimens," *The Journal of the American Medical Association*, vol. 313, no. 11, pp. 1122–1132, 2015.

[55] D. B. Nagarkar, E. Mercan, D. L. Weaver, T. T. Brunyé, P. A. Carney, M. H. Rendi, A. H. Beck, P. D. Frederick, L. G. Shapiro, and J. G. Elmore, "Region of interest identification and diagnostic agreement in breast pathology," *Modern Pathology*, 2016.

[56] E. A. Krupinski, A. R. Graham, and R. S. Weinstein, "Characterizing the development of visual search expertise in pathology residents viewing whole slide images," *Human Pathology*, vol. 44, no. 3, pp. 357–364, 2013.

[57] A. Vedaldi and K. Lenc, "Matconvnet – convolutional neural networks for matlab," 2015. http://www.vlfeat.org/matconvnet.

[58] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, 1998.