Automated Histopathological Analyses at Scale

by

Mrinal Mohit

B. Tech., Indian Institute of Technology Kharagpur (2015)

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of

Master of Science in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2017

Signature redacted

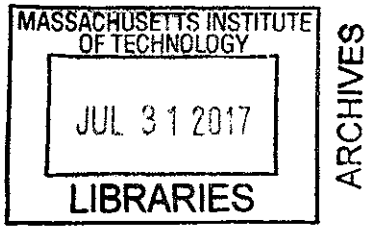
Author .....
Program in Media Arts and Sciences
May 12, 2017

Signature redacted

Certified by .....
Ramesh Raskar
Associate Professor
Program in Media Arts and Sciences
Thesis Supervisor

Signature redacted

Accepted by .....
Pattie Maes
Academic Head
Program in Media Arts and Sciences



Automated Histopathological Analyses at Scale

by

Mrinal Mohit

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
on May 12, 2017, in partial fulfillment of the
requirements for the degree of
Master of Science in Media Arts and Sciences

Abstract

Histopathology is the microscopic examination of processed human tissues to diagnose conditions like cancer, tuberculosis, anemia and myocardial infractions. The diagnostic procedure is, however, very tedious, time-consuming and prone to misinterpretation. It also requires highly trained pathologists to operate, making it unsuitable for large-scale screening in resource-constrained settings, where experts are scarce and expensive.

In this thesis, we present a software system for automated screening, backed by deep learning algorithms. This cost-effective, easily-scalable solution can be operated by minimally trained health workers and would extend the reach of histopathological analyses to settings such as rural villages, mass-screening camps and mobile health clinics. With metastatic breast cancer as our primary case study, we describe how the system could be used to test for the presence of a tumor, determine the precise location of a lesion, as well as the severity stage of a patient. We examine how the algorithms are combined into an end-to-end pipeline for utilization by hospitals, doctors and clinicians on a Software as a Service (SaaS) model. Finally, we discuss potential deployment strategies for the technology, as well an analysis of the market and distribution chain in the specific case of the current Indian healthcare ecosystem.

Thesis Supervisor: Ramesh Raskar
Title: Associate Professor, Program in Media Arts and Sciences

Automated Histopathological Analyses at Scale

by

Mrinal Mohit

The following people served as readers for this thesis:

Signature redacted

Thesis Reader .....

Edward Boyden
Associate Professor of Media Arts and Sciences
Program in Media Arts and Sciences

Signature redacted

Thesis Reader .....

Sepandar Kamvar
Associate Professor of Media Arts and Sciences
Program in Media Arts and Sciences

Acknowledgements

I am grateful to my advisor, Prof. Ramesh Raskar, for his continued support, guidance and mentorship. His suggestions, advice and feedback on the project have been invaluable for the work presented in this thesis.

I also thank the Camera Culture group at the MIT Media Lab, especially Anshuman Das, Tristan Swedish and Otkrist Gupta, for their invigorating discussions and helpful inputs.

This work uses publicly available datasets graciously collected and hosted by the Radboud University Medical Center and Utrecht University Medical Center, Netherlands.

This work was funded by the MIT Tata Center for Technology and Design and the MIT Media Lab consortium.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The field of pathology deals with providing clinical disease diagnoses to inform patient treatment and management. Accurate and reproducible pathology diagnoses are essential for precision and targeted medicine. The primary tool used by pathologists has been the microscope, followed by a qualitative visual analysis [1] [39]. However, this has been limited by lack of standardized agreement between pathologists, as well as diagnostic errors arising from the task of manually analyzing gigapixel images for millions of cells across hundreds of slides on a daily basis [45] [49] [8], with inter-rater agreement sometimes as low as 48%, and only 70% on average. [17].

1.1 Motivation

As a case study for a particular pathology, we consider detecting metastatic breast cancer in digital whole slide images of lymph node biopsies. (Figure 1-1) According to the American Joint Committee on Cancer [16], a patient with sentinel lymph nodes diagnosed to be positive for breast cancer is assigned a more severe pathological

stage, with more aggressive interventions (like auxiliary lymph node dissection) than a patient with a negative result [43] [44] [3].

The manual visual analysis of sentinel lymph node slides is tedious, repetitive and laborious, especially for negative cases or cases with only very small affected regions. This often results in diagnostic errors and false negatives. Chemical alternatives for analysis like immunohistochemistry are limited by increased cost, increased processing and increased number of slides required, with little improvement in accuracy. [14].

The development of effective and cost-efficient computer-assisted image analysis systems for lymph node evaluation remains an area of active research, as a high-performing system which could improve accuracy while reducing cognitive load would be highly valuable. [34]

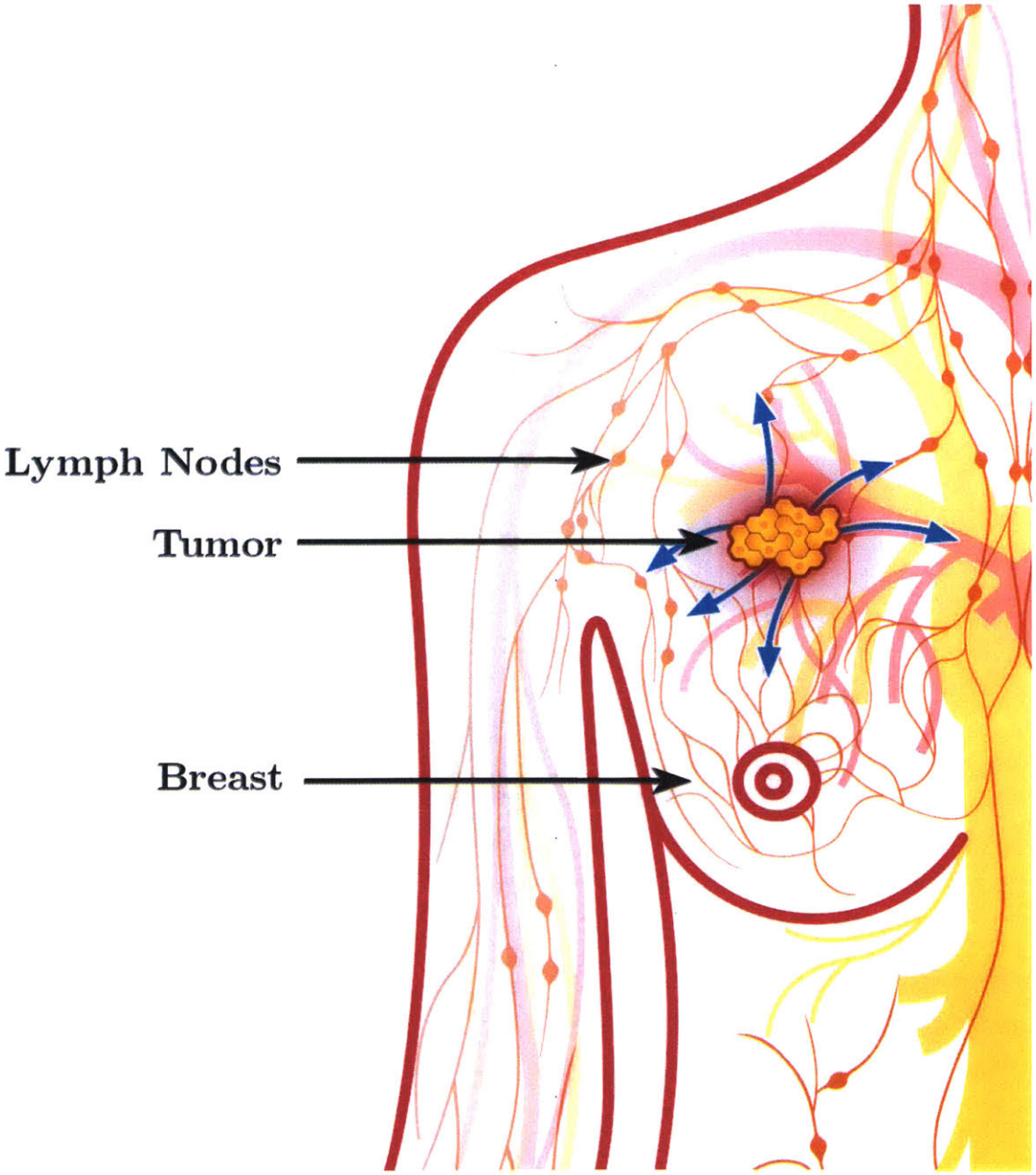
In this thesis, we present a deep learning based approach for histopathological analyses, specifically identification and localization of breast cancer metastases from whole slide images of sentinel lymph nodes. We further study how these analyses from multiple slides for a patient can be synthesized into an estimate of the clinical stage of the patient, and how these methods can be evaluated and compared to the state-of-the-art.

Additionally, we further present an analysis of how the algorithms can be combined into an end-to-end system for deployment, and what those deployment opportunities and challenges could be in the specific case of the Indian healthcare system.

1.2 Contributions

This thesis makes important contributions in showing how deep learning solutions can be used to automate histopathological analyses at scale, lowering cost and extending

Figure 1-1: Lymph nodes are small glands that filter lymph, a colorless fluid containing white blood cells that bathes the tissues and drains through the lymphatic system into the bloodstream. Metastatic involvement of lymph nodes is one of the most important prognostic factors in breast cancer. They are surgically removed and examined microscopically, as in the whole-slide images analyzed in the following chapters. (Illustration adapted from the National Breast Cancer Foundation [21])



reach in resource-constrained settings. In the subsequent chapters, we describe in detail the following -

- **Algorithms for identifying and classifying whole slide images of sentinel lymph nodes for metastatic breast cancer.**
  - A split-and-recombine strategy to train convolutional neural networks and employ them for localized predictive inference.
  - An aggregation framework for post-processing likelihood maps into whole-slide classification results.
- **Algorithms for precise automated localization of suspected tumor sites in whole slide images.**
  - A two-stage deep thresholding model which uses connected components in the likelihood maps to identify tumor locations.
  - A region proposal network which learns to identify areas of potentially high information to guide search for tumor boundaries.
- **Algorithms for synthesizing evaluations of multiple slides from different lymph nodes to characterize the clinical stage of a patient.**
  - A rule-based learner which incorporates guidelines followed by pathologists to rate clinical stage.
  - A joint learning model for both likelihood and stage, which is able to encode information not captured by the guidelines.
- **Analyses of deployment strategies for the software solution in India.**
  - A study of the key customer segments and stakeholders and promising business models in the Indian healthcare industry.

– An analysis of factors required to set up a sustainable platform and a chain for distribution.

Chapter 2

Background and Related Work

The conventional qualitative visual analysis for histopathological images relies on the digital imaging technologies employed to capture the relevant tissue image, and the expertise of the trained pathologist. The work presented in this thesis aims to present a number of methods to augment or enhance the latter component with computer vision and deep learning techniques. In this respect, this work draws from three broad areas - digital pathology, traditional computer vision and recent advances in deep learning. Following are key developments in each field which form the basis of this work.

2.1 Digital Pathology

Tissue samples for pathological analyses are usually collected via surgery or via biopsy, and need to be further processed in order to make glass slides which hold histological sections of the thickness of a few micrometers. This processing includes fixation, embedding, cutting and staining. During the preparation of histological slides, different

stains can be used for various purposes, but the hematoxylin and eosin (H&E) stain is most widely used; it induces sharp blue/pink contrasts across various sub-cellular structures and it is applied across many different tissue types.

Digital Imaging for pathology has gained acceptance in the last decade as the de facto representation modality, for easy access and archiving [23]. Whole-slide scanners are used to digitize glass slides at high resolution (up to 160nm per pixel). The availability of whole-slide images (WSI) has sparked the interest of the medical image analysis community, resulting in increasing numbers of publications on histopathological image analysis.

These whole-slide images are generally stored in multi-resolution TIFF pyramid files. Image files contain multiple down-sampled versions of the original image. Each image in the pyramid is stored as a series of tiles, to facilitate rapid retrieval of subregions of the image.

A typical whole-slide image is approximately 200,000 x 100,000 pixels on the highest resolution level with 3 byte RGB pixel format. This translates into over 56 GB of uncompressed pixel data from a single level. This has resulted in a need for clever techniques to efficiently handle and process such amounts of data.

2.2 Classical Computer Vision

Over the past decades, there has been an increased interest in computational methods for assistance in analysis of microscopy images in pathology. [27]

Historically, approaches to histopathological image analysis in digital pathology have focused primarily on low-level image analysis tasks (e.g., color normalization, nuclear segmentation, and feature extraction), followed by construction of classification

models using classical machine learning methods, including: logisitic regression, support vector machines, and random forests. Typically, these algorithms take as input relatively small sets of image features (10-100) [27] [32]. Further building on this framework, approaches have been developed for the automated extraction of moderately high dimensional sets of image features (1000s) from histopathological images followed by the construction of relatively simple, linear classification models using methods designed for dimensionality reduction, such as sparse regression [5].

2.3 Deep Learning

Since 2012, deep Convolution Neural Networks have significantly boosted accuracy on a wide range of computer vision tasks, including image recognition [38] [31] [52], object detection [25] [24] [50] and semantic segmentation [42]. They have also been employed productively in healthcare applications, notably detection of diabetic retinopathy from retinal fundus photographs [26], and dermatologist-level detection of melanoma from skin photographs [20].

In the context of histopathology, there have been studies in applying the above techniques to detect prostate cancer biopsies [41] and to segment epithelium, tubules, lymphocytes, mitosis and lymphoma [35]. It has been demonstrated that CNNs achieved higher F1 score and balanced accuracy in detecting invasive ductal carcinoma.[13]. Mitosis detection competitions at ICPR 2012 [10] and AMIDA 2013 [61] have been won by teams which used CNNs. Predicting prognosis in non-small cell lung cancer using machine learning was described in [63]. In contrast to the types of machine learning approaches historically used in digital pathology, in deep learning-based approaches there tend to be no discrete human-directed steps for object detection, object segmentation, and feature extraction. Instead, the deep learning algorithms take as input only the images and the image labels (e.g., 1 or 0) and learn a very high-

dimensional and complex set of model parameters with supervision coming only from the image labels.

It has been hypothesized that deep learning techniques could increase the sensitivity, speed and consistency of metastasis detection, but so far no systems have been able to deliver on those fronts to a clinically acceptable standard. [41]

Chapter 3

Algorithms for Lesion Classification

Classifying whole-slide images of sentinel lymph node tissues extracted from suspected patients into normal or tumor is an important, and often critical, prognostic step for the pathologist. In this chapter, we describe the dataset we used to train our models on, the evaluation metrics to judge performance, the classification pipeline, and close with a discussion of the results.

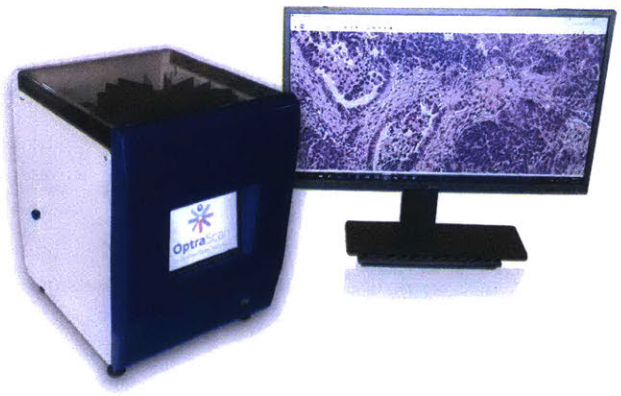
3.1 Dataset

Whole-slide scanners (Figure 3-1) were used to collect images of sentinel lymph node tissues after fixation, embedding, cutting and staining by hematoxylin and eosin (H&E). These scanners can reach resolutions of upto 160 nanometer per pixel, and output images at multiple scanning resolutions in a pyramid structure (Figure 3-2). The data was collected by the Diagnostic Image Analysis Group (DIAG) and

Department of Pathology of the Radboud University Medical Center (Radboudumc) in Nijmegen, The Netherlands, and made available publicly in the Camelyon challenge [22].

The data consists of a total of 400 whole-slide images of sentinel lymph nodes, split into a training set of 270 slides (160 normal, 110 tumor), and a test set of 130 slides. The ground truth data for the training slides consists of a pathologists delineation of regions of metastatic cancer on WSIs of sentinel lymph nodes. The data was provided in two formats: XML files containing vertices of the annotated contours of the locations of cancer metastases and WSI binary masks indicating the location of the cancer metastasis.

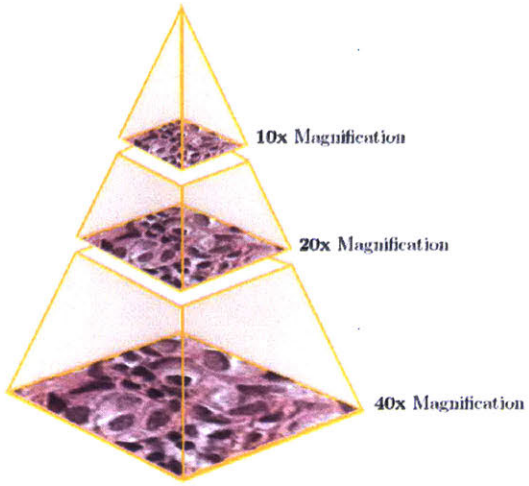
Figure 3-1: A typical whole-slide scanner system (courtesy of OptraScan [47]). These solutions can scan over 120 slides at once, with the option of 10x, 20x or 40x magnification. The output is usually in multi-resolution TIFF image pyramids and of resolutions in the order of 200,000 x 100,000 pixels at highest magnification.

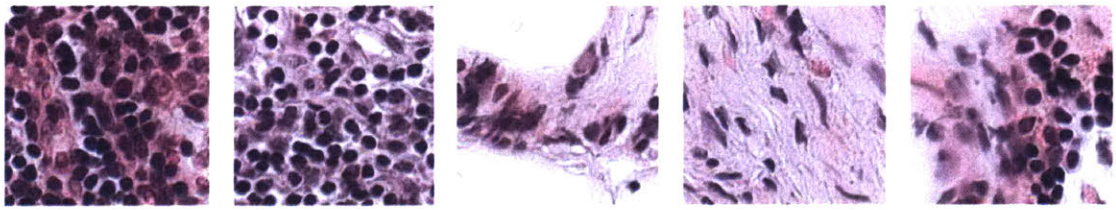


3.2 Evaluation Metrics

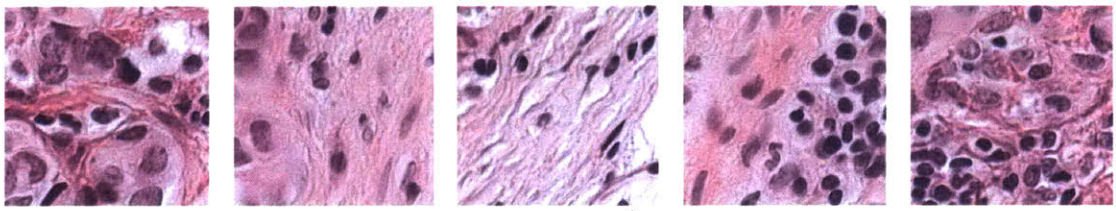
The whole slide classification task between metastasis and normal slides is assessed by the means of the area under (AUC) the receiver operating characteristic (ROC) curve, which is a graphical plot that illustrates the performance of a binary classifier system

Figure 3-2: Structure of a TIFF image pyramid file - the output of a whole slide scanner. It usually is a 3 channel image over multiple optical resolutions (the highest magnification can reach 160 nm per pixel), for easy retrieval of relevant subregions.





(a) Examples of normal patches



(b) Examples of tumorous patches

Figure 3-3: Slide samples at highest magnification

as its discrimination threshold is varied. It is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity or recall, while the false-positive rate is also known as the fall-out or probability of false alarm and can be calculated as (1 specificity). The ROC curve is thus the sensitivity as a function of fall-out [4]. For each image in the test set, the algorithm ultimately provides the probability of that image to contain metastasis.

3.3 Classification Pipeline

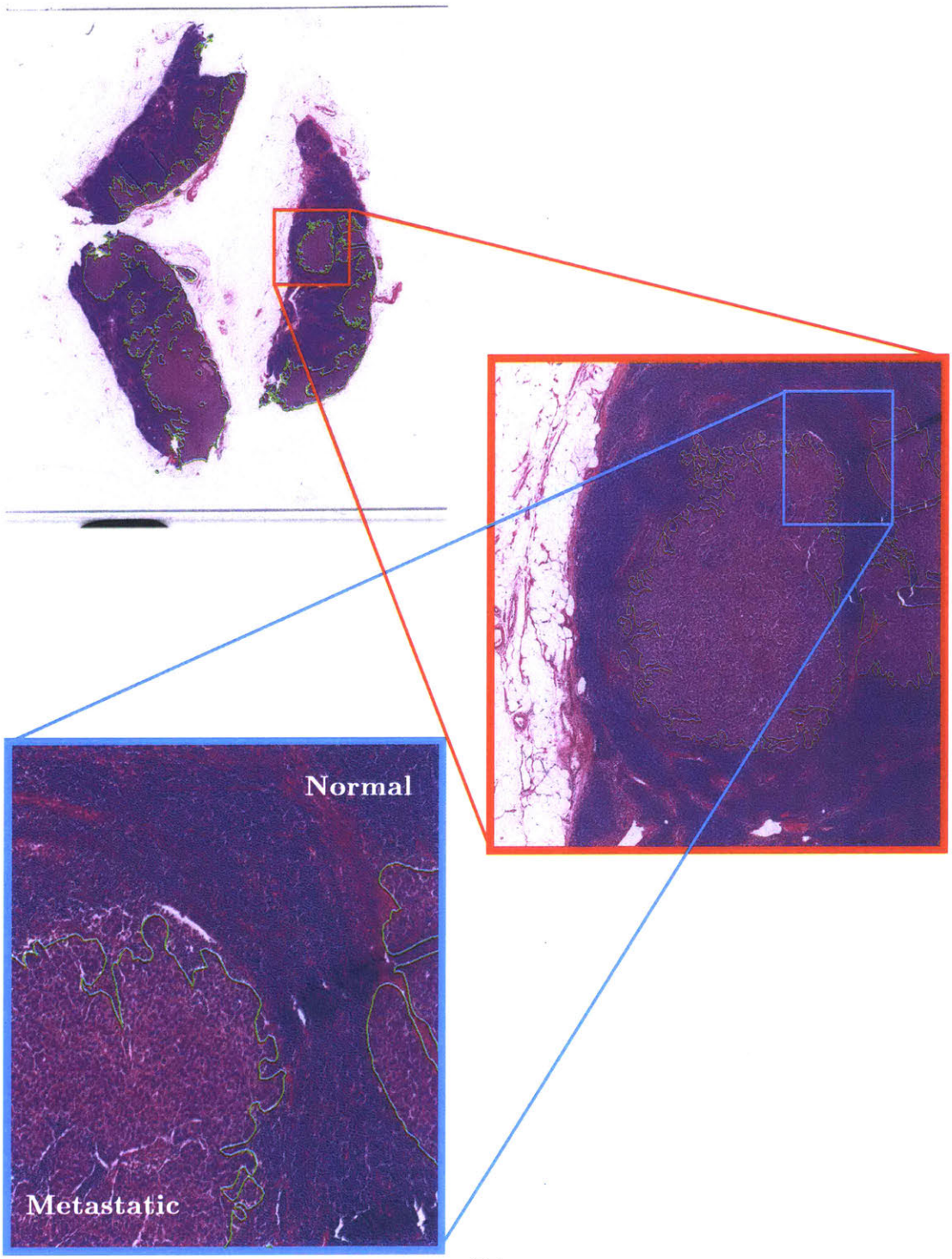
Because of the extremely large size of each slide and the limited number of slides, we divide each whole-slide image into small square patches and train a deep convolutional neural network to make patch-level predictions to discriminate tumor-patches from normal-patches. We then aggregate the patch-level predictions to create tumor likelihood heatmaps and perform post-processing over these heatmaps to make predictions for the slide-based classification task (and later the tumor-localization task).

3.3.1 Pre-processing

The whole slide images invariably contain a lot of negative space (80%), which greatly increases computation time (with each image being 56 GB uncompressed). To reduce the required compute and to focus on the regions most likely to contain cancer metastasis, we identify the tissue the image. This is achieved by an unsupervised thresholding process -

- The image is transformed from RGB to HSV cylindrical-coordinate color space to reduce dependence of the procedure on lighting and contrast. The color (Hue, Saturation channels) is separated from the lighting (Value channel). [2]

Figure 3-4: Example of a whole-slide image with macro-metastases. The structures outlined in green represent the tumors in the sample (according to the pathologist-labeled ground truth annotations). At high magnification, the differences in local texture between the normal and metastatic regions. However, as later discussed, these differences can often be very subtle.

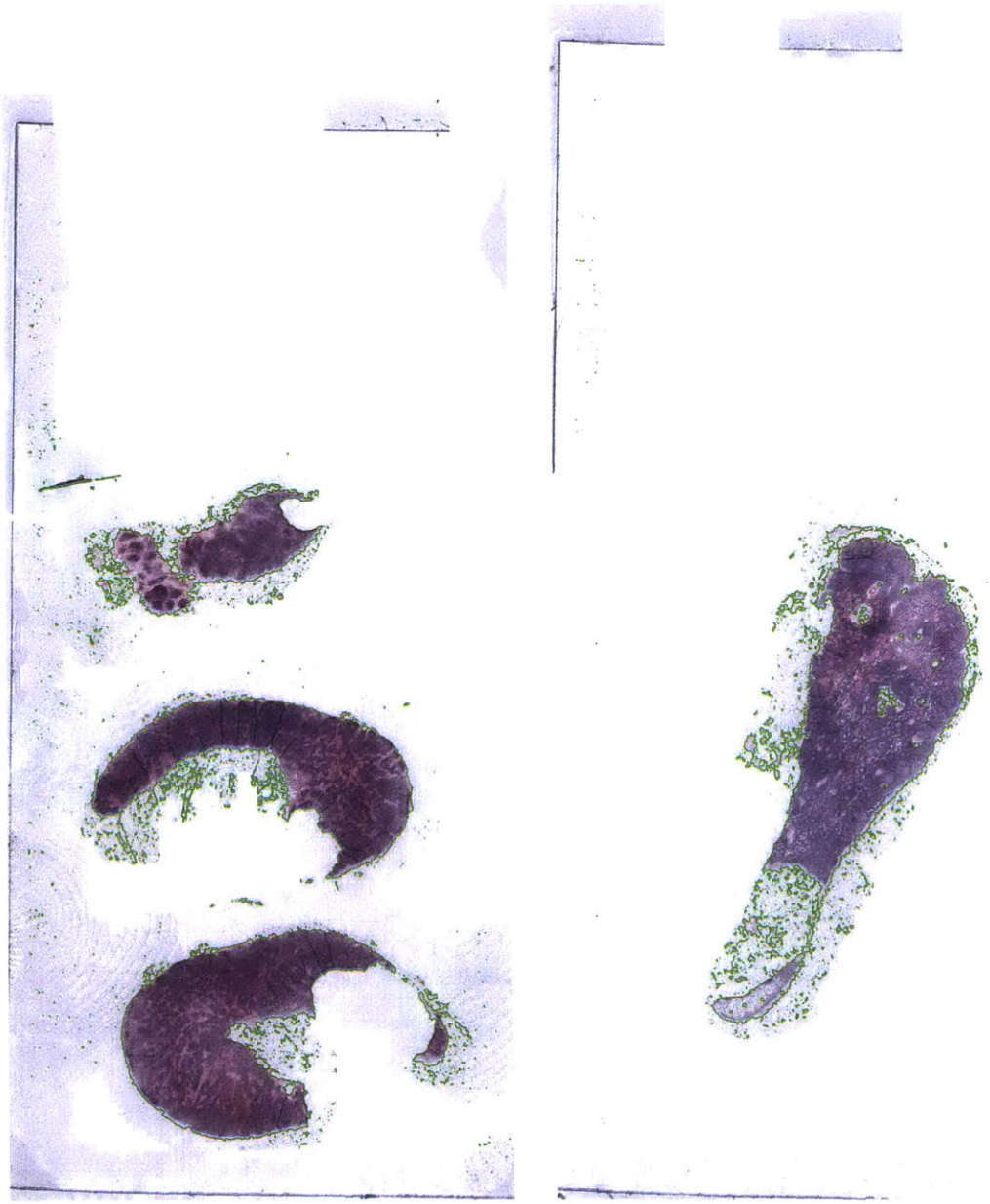


- Optimal threshold values are computed for each channel with Otsu's algorithm for binary segmentation [48], which minimizes intra-class variance or equivalently, maximizes inter-class variance.
- A union is performed for the mask for H and S channels, and then its intersection with the original whole-slide image is taken as the new data source.

The reduction in area to consider is shown in Figure 3-5, with the threshold boundaries identified in green. On an average 81% of the data per whole-slide image is removed with this process, making subsequent steps computationally tractable.

We also perform stain specific standardization of the whole-slide images, analogous to that described in [6], to make the process invariant to most inconsistencies in the composition of the stain and the process of applying it. In order, we

1. Transform to the Hue-Saturation-Density space
2. Extraction samples for the hematoxylin, eosin and background classes from the WSI and derive the chromatic and density distributions of these classes.
3. Transform the 2D chromatic distribution for each dye class to match the chromatic distribution of the corresponding class from a template slide.
4. Transform the density distribution for each dye class to match the density distribution of the corresponding class from a template slide.
5. Weigh the contribution of stains for each pixel and obtain final chromatic and density transformations.
6. Apply inverse HSD transform.



(a) (b)

Figure 3-5: Results of background removal. The green lines delineate the regions ultimately processed by the subsequent steps. On an average, 81% of the data is culled.

3.3.2 Data Augmentation

In order to further make the classifier robust to variance in orientation, stain color and other small perturbances, we augment our dataset with artificially generated samples

- We perform random translations by cropping a 224x224 subregion from the patch.
- We perform random horizontal flips and rotations by 90 degree multiples.
- We perform random aspect ratio modifications in the range 3/4 to 4/3.
- We perform color perturbation by adding to the RGB channels multiples of their principal components, scaled by the eigenvalues times a normal random variable drawn from $N(0,1)$. i.e. to each RGB image pixel $I_{xy} = [{I_{xy}}^R, {I_{xy}}^G, {I_{xy}}^B]^T$, we add the quantity

  $$[p_1, p_2, p_3][\alpha_1\gamma_1, \alpha_2\gamma_2, \alpha_3\gamma_3],$$

  where p_i and γ_i are ith eigenvector and eigenvalue of the 3×3 covariance matrix of RGB pixel values, respectively, and α_i is the aforementioned random variable. This scheme approximately captures an important property of natural images, namely, that object identity is invariant to changes in the intensity and color of the illumination.
- We add jitter to the patch extraction process so that each patch has a small x,y offset of up to 8 pixels.

The modifications are performed on-the-fly during training so that the transformed images need not be stored on disk, and sequenced on the CPU while the GPU is processing the last batch, severely reducing computational needs.

3.3.3 Patch-based Classification

The patch-based classification stage takes as input whole slide images (of resolution in the order of 200,000 x 100,000 pixels) and the ground truth image annotation delineating the locations of regions of each WSI containing metastatic cancer. We randomly extract millions of small (256 x 256 pixels) positive and negative patches from the set of training WSIs. If the small patch is located in a tumor region, it is a tumor / positive patch and labeled with 1, otherwise, it is a normal / negative patch and labeled with 0. On an average, we extract, without overlap, 1000 positive and 1000 negative samples from each tumor image (and only the latter from a normal images). This results in a total of 110,000 positive and 270,000 negative data points, representative samples of which are illustrated in Figure 3-4.

Following selection of positive and negative training examples, we use a deep neural network to train a supervised classification model to discriminate between these two classes of patches. Deep learning enables a hierarchy of processing layers to automatically learn representation of data with an increasing level of abstraction. It discovers intricate structure in large data sets by using the backpropagation algorithm to suggest how the computer should change its internal parameters that are used to compute the representation in each layer from the representation in the previous layer. Deep convolutional networks have recently brought about breakthroughs in processing images, video, speech and audio. [40]

This is in sharp contrast to conventional machine-learning techniques which require careful engineering and considerable domain expertise to design a feature extractor that transformed the raw data into a suitable internal representation or feature vector from which the learning subsystem could detect patterns in the input. The key disadvantage of the using classical linear classifiers (which compute a weighted sum of the feature vector and compare it to a threshold) in the network is their limitation of

only being able to carve the input space into simple regions, i.e. half-spaces separated by a hyperplane. This works well for certain kinds of data but if applied to image recognition, it results in insufficient invariance to position, orientation or illumination.

During the learning, the network (examples of the structure are described later), is shown an image patch and it produces an output probability for it being a tumor. The training dataset has images $x_i \in R^D$, each associated with a label y_i. Here $i = 1 \ldots N$ and $y_i \in 1 \ldots K$. That is, we have N examples (each with a dimensionality D) and K distinct categories. The score function $f : R^D \mapsto R^K$ maps the raw image pixels to class scores. An integral part of the multi-layered score function is regular non-linearities, which distinguish it from a linear classifier.

This score is unlikely to be correct before training. We compute a loss function that measures the error / distance between the output score and the desired score. We use the cross-entropy loss with weight decay, with each training example additively contributing $L_i = f_{y_i} + log \sum_j e^{f_j}$, where f refers to the vector of class scores. Minimizing this loss is equivalent to minimizing the Kullback-Leibler divergence between the estimated class probability distribution and the underlying ground truth.

The network then adjusts its parameters (or weights) to reduce this error. The number of these weights are in the order of millions, and typically require millions of training examples to converge to a high enough accuracy. The adjustment of the weights is performed by computing the gradient of the error with respect to each weight, and then the weight is adjusted in the opposite direction, scaled by the learning rate l. If performed over multiple layers, this results in the widely used backpropagation algorithm [30].

However, since computing the objective function over all the training examples could be very slow, Stochastic Gradient Descent (SGD) is used as an approximation. It divides the training data into a mini-batch of examples over which the gradient is

computed. Increasing the mini-batch size trades off frequency of update for gradient accuracy and smoothness [7], but in practice having hundreds of examples in a batch is sufficient for a reasonable gradient estimate. We initialize weights to small random numbers, accounting for the parameter count [28] and force activations throughout a network to take on a unit gaussian distribution at the beginning of the training (Batch Normalization [31]), along with L2 weight decay and dropout [58]). We further use the Adam optimization technique [36], modeled as

$$m = \beta_1 * m + (1 - \beta_1) * dx$$
$$v = \beta_2 * v + (1 - \beta_2) * dx^2$$
$$x = x - \frac{l * m}{\sqrt{v} + \epsilon}$$

where x is the gradient, m is the smoothened gradient, l is the learning rate, and β_1, β_2 and ϵ are hyperparameters. The trained network is finally run on a hitherto unobserved test set (or in this case, a 15% validation set).

Convolutional architectures make the explicit assumption that the inputs are images, which allows us to encode certain properties into the architecture. These then make the forward function more efficient to implement and vastly reduce the amount of parameters in the network. The ones detailed in Figure 3-6 are composed of the following layers

- Input - holds the raw pixel values of the image, in this case an image of width 256, height 256, and with three color channels R,G,B.
- Convolution - computes the output of neurons that are connected to local regions in the input, each computing a dot product between their weights and a small region they are connected to in the input volume.
- ReLU - applies an elementwise non-linear activation function, such as $max(0, x)$

- MaxPool - performs a downsampling operation along the spatial dimensions (width, height)
- FC (fully-connected) - computes the final class scores.

We evaluate three deep learning network architectures for performing classification between the normal and the tumorous patches - VGG [55], Inception [59] and Residual Networks [29]. The newly-introduced ResNets perform better than the previous state-of-the-art, as shown in Table 3.1

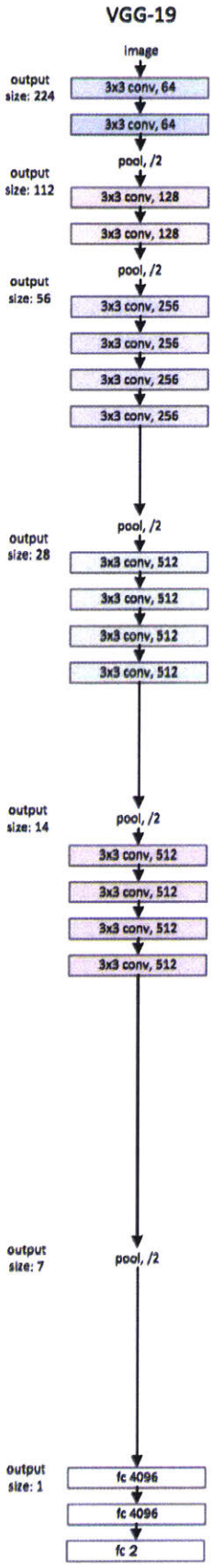
Network Architecture	**Accuracy**
VGG-19 [55]	97.7%
Inception [59]	98.2%
ResNet-34 [29]	98.8%

Table 3.1: Performance of deep learning architectures

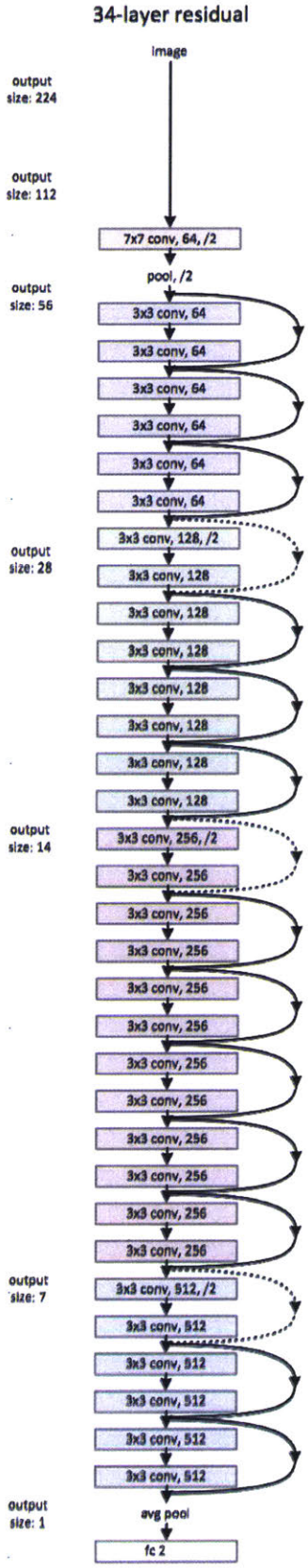
To bootstrap and learn from false positives, we extract additional negative samples from the non-tumor regions (200 per WSI) and retrain the model with the enhanced dataset.

Finally, we embed all the prediction (at stride 256) results into a likelihood (heatmap) image. Each pixel contains a value between 0 and 1, indicating the probability that the pixel contains tumor. Figure 3-8 and 3-9 present results of the patch classification on one of the whole-slide samples (with the overlaid annotated tumor) and the corresponding output likelihood map.

All experiments were performed with Torch [12] running on a server with 2x 8-core Intel Xeon E5-2620 @2.1 GHz, 128 GB DDR4 RAM, 2 TB SSD and 8x GTX 1080 8 GB. 8 replicas were run the GPUs with asynchronous gradient updates and batch size of 64 per replica. We used Adam [36] with ϵ of 10^{-8}, β_1 of 0.9 and β_2 of 0.999. The initial learning rate was 0.07, with a decay of 0.5 every 1 million examples.



(a) VGG-19



(b) ResNet-34

3.3.4 Post-processing of likelihood maps

We take as input a heatmap for each whole-slide image and produce as output a single probability of tumor for the entire WSI. We extract 28 geometrical and morphological features (skimage's regionprops [60]) from each heatmap, including the percentage of tumor region over the whole tissue region, the area ratio between tumor region and the minimum surrounding convex region, the average prediction values, and the longest axis of the tumor region. We compute these features over tumor probability heatmaps across all training cases, and we build a 50-tree random forest classifier to discriminate the WSIs with metastases from the negative WSIs.

On analyzing the relative importance of each feature, the following proved to be the most discriminatory -

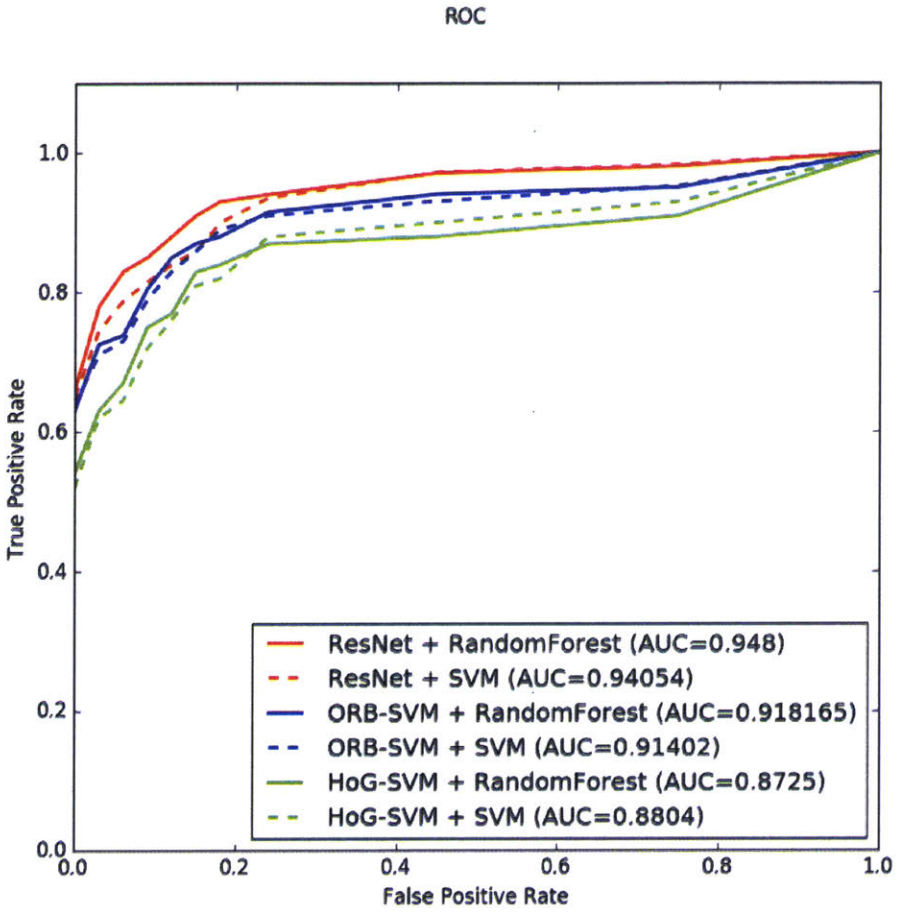
1. With threshold = 0.5, the longest axis in the largest tumor region
2. With threshold = 0.5, ratio of pixels in the region to pixels in the total bounding box (extent)
3. Eccentricity of the ellipse that has the same secondmoments as the region. (eccentricity)
4. Ratio of tumor region with threshold = 0.9 to the tissue region
5. With threshold = 0.5, the area of largest tumor region

3.4 Results and Discussion

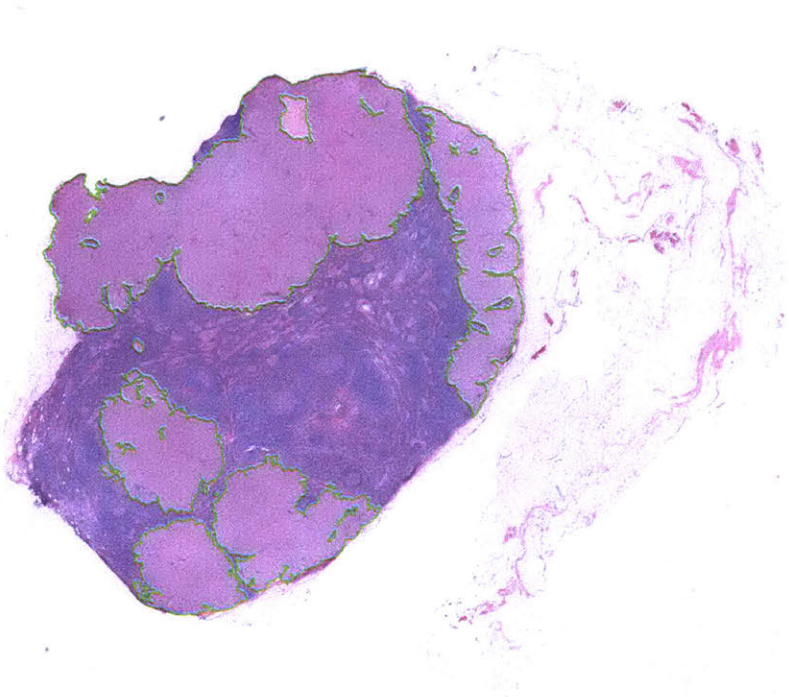
We compared our classification receiver-operator characteristics to those previously considered-state-of-the-art. We test it against the performance obtained from support

vector machines trained on Histogram of Oriented Gradients (HoG) [15] features and Oriented FAST and Robust BRIEF (ORB) [51] features of the input image data. The ROC curves in Figure 3-7 show how using a deep network boosted the AUC significantly.

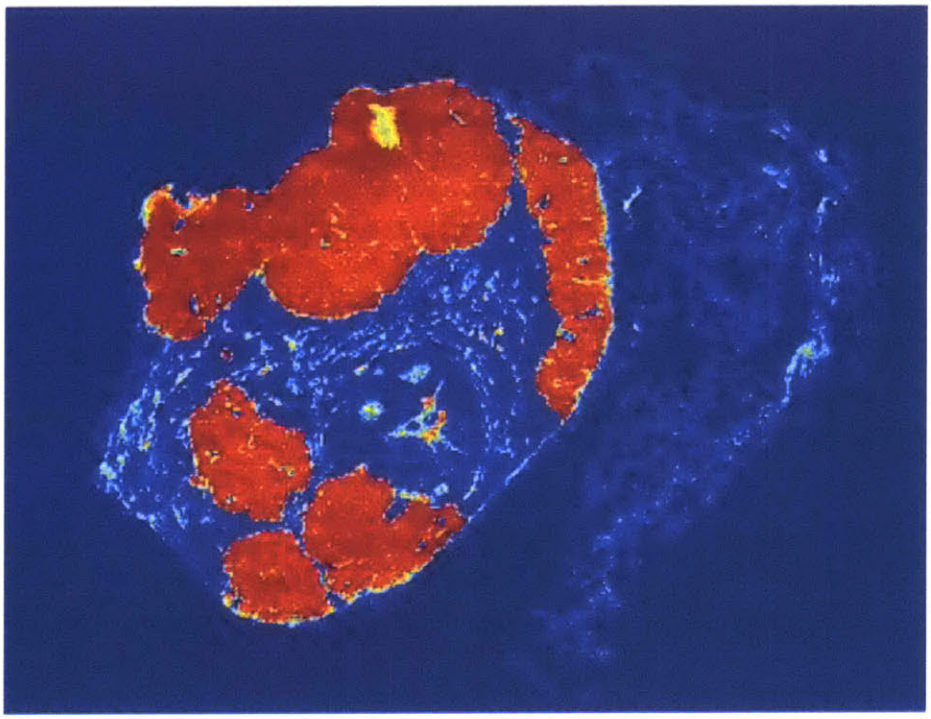
Figure 3-7: Performance Characteristics on the Lesion Classification Task



We also note that swapping out a random forest classifier in the post-processing step works better than a support vector machine, although not by much. We note that this metric is overall a challenging one because of the potential for false positives increases dramatically when a high level of patch-level predictions are obtained per slide. In contrast, human pathologist rarely demonstrate false positives in their analysis [35].

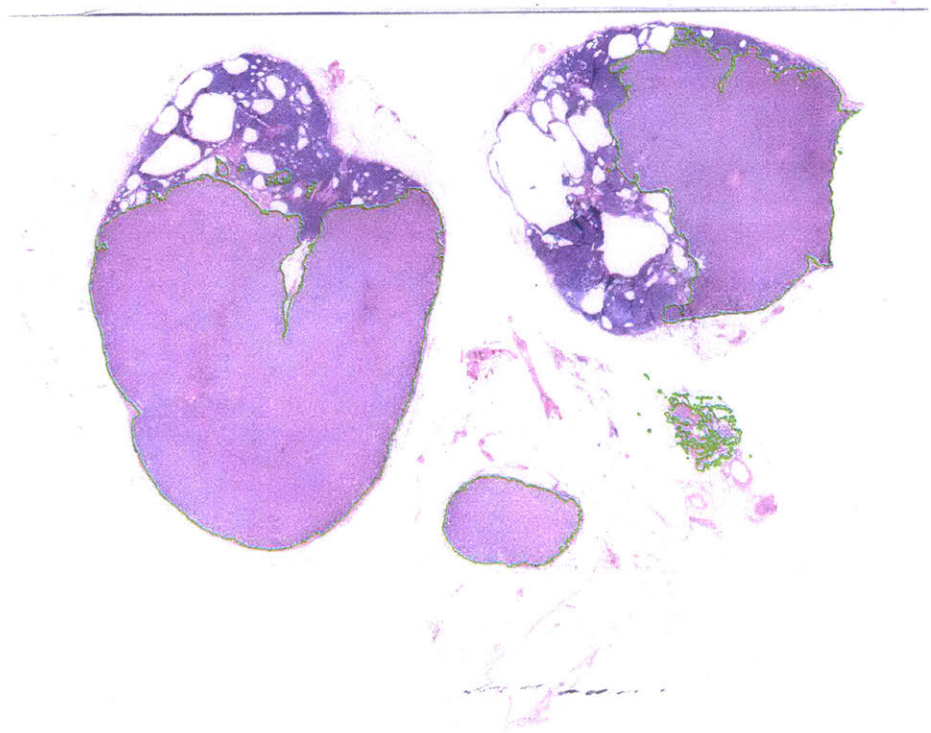


(a) Whole-slide image with annotated tumors in green

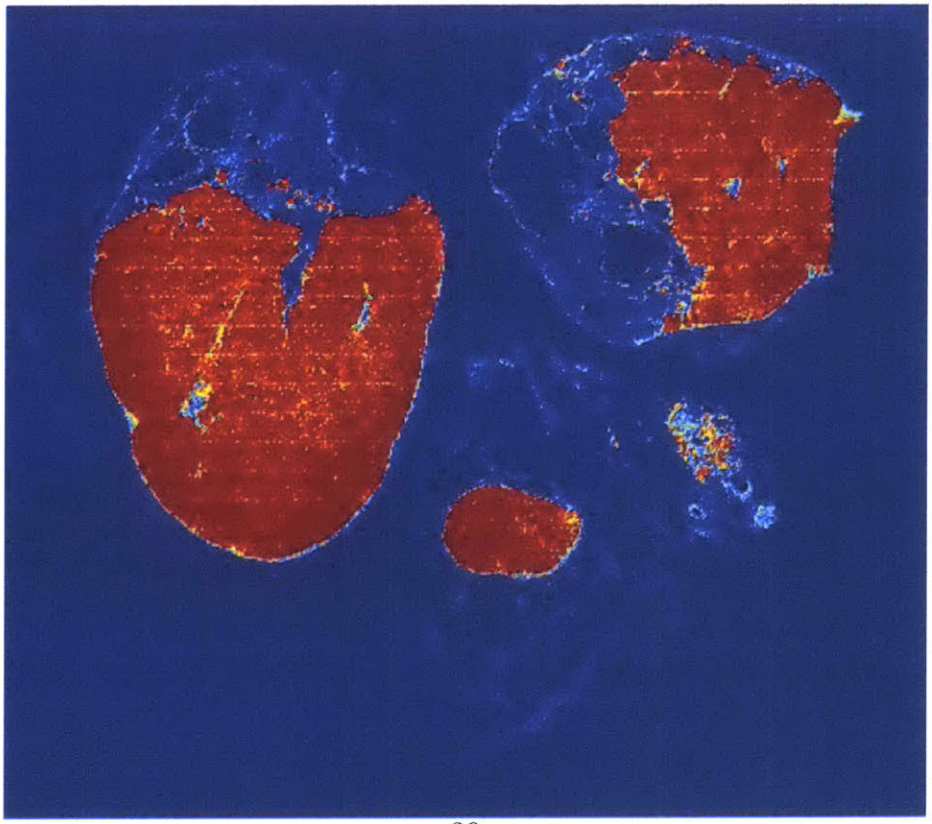


(b) Tumor likelihood map generated by the patch classification stage. Since each pixel corresponds to a 256x256 patch in the original slide, the resolution is 65,000x lower.

Figure 3-8: Patch classification result sample 1.



(a) Whole-slide image with annotated tumors in green



(b) Tumor likelihood map generated by the patch classification stage.

Figure 3-9: Patch classification result sample 2.

Chapter 4

Algorithms for Lesion Localization

The precise localization of lesions in a whole-slide image, especially of those that are very small in size is an important prognostic step, since the geometrical characteristics of the abnormality - like the size, eccentricity, location with respect to other cells etc., inform the diagnosis. It also is extremely tedious and laborious for the pathologist and is thus prone to errors. To explore a variety of methods to approach this task, we use the same dataset as described in the previous chapter.

4.1 Evaluation Metrics

The lesion localization performance is summarized using the Free Response Operating Characteristic (FROC) curve. This is similar to ROC analysis, except that the false positive rate on the x-axis is replaced by the average number of false positives per image. A true positive is considered if the location of the detected region is within the annotated ground truth lesion. If there are multiple findings for a single ground truth region, they are counted as a single true positive finding and none of them are

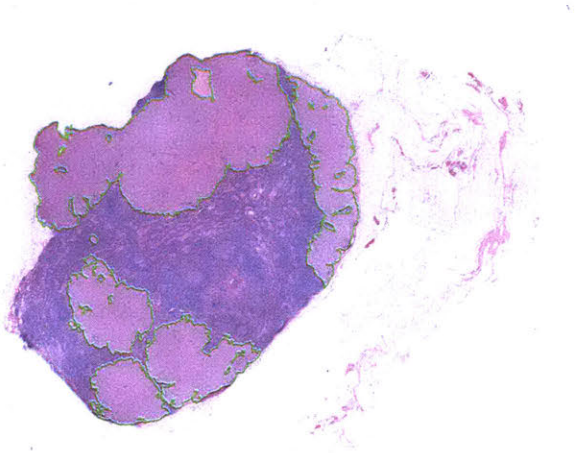
counted as false positive. All detections that are not within a specific distance from the ground truth annotations are counted as false positives. The final score is defined as the average sensitivity at 6 predefined false positive rates: 1/4, 1/2, 1, 2, 4, and 8 FPs per whole slide image.

For each tumor region detected in the whole-slide image (WSI), the algorithm provides the X and Y coordinates of the detected region together with a confidence score representing the probability of the detected region to be tumorous.

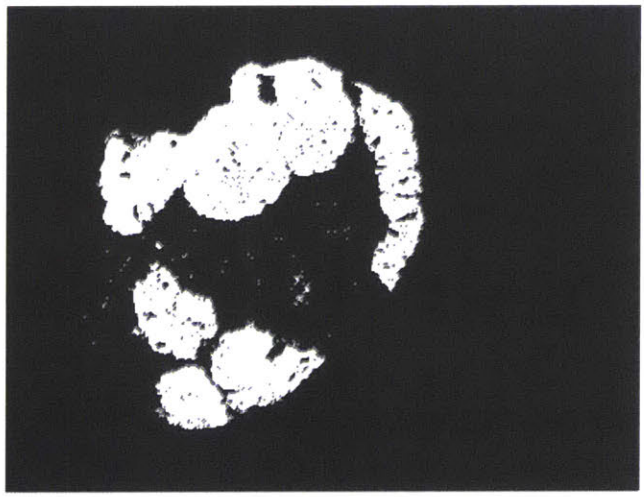
4.2 Two-stage thresholding model

We aimed to identify all cancer lesions within each whole-slide image with as few false positives as possible. We perform the pre-processing and patch classification steps described in Chapter 3, but proceed differently for the post-processing steps.

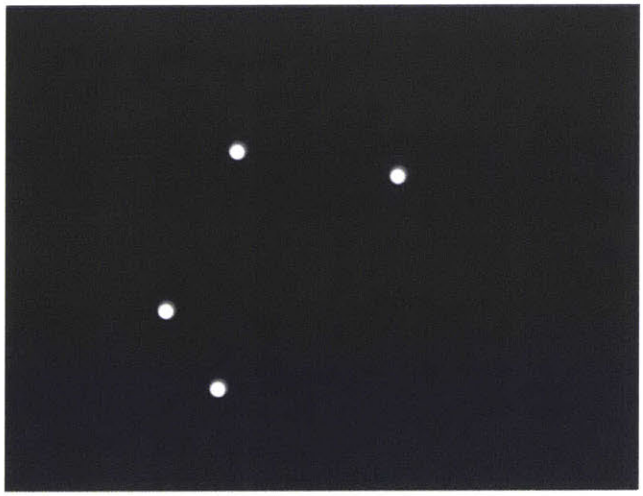
- We first train a deep model M_A using our initial training dataset described in Chapter 3.
- We then train a second deep model M_B with a training set that is enhanced with a higher representation of negative regions adjacent to tumors. This model M_B produces fewer false positives than M_A but has reduced sensitivity.
- We threshold the heatmap produced from M_A at T_A, which creates a binary mask.
- We then identify connected components within the tumor binary mask, and we use the central point as the tumor location for each connected component.
- To estimate the probability of tumor at each of these (x, y) locations, we take the average of the tumor probability predictions generated by M_A and M_B across each connected component.



(a) Whole-slide image with annotated tumors in green



(b) Tumor likelihood map thresholded at T_A. Since each pixel corresponds to a 256x256 patch in the original slide, the resolution is 65,000x lower.



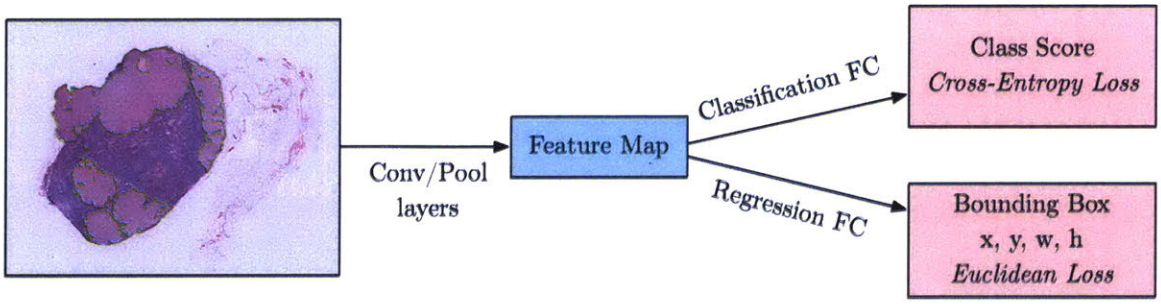
(c) Centres of connected components in the thresholded map

Figure 4-1: Two-stage thresholding result sample.

4.3 Localization as Regression

To explore if we can directly train the convolution network to predict boundaries of the tumor, we pose it as a regression problem for the location and size of their bounding boxes, which are then accumulated rather than suppressed in order to increase detection confidence. We apply this ConvNet at multiple locations in a sliding window fashion over multiple scales (4 here), which are then merged. The modified scheme for the network is shown in Figure 4-2. This method also avoids using background samples resulting in much faster training and side-stepping the bootstrapping process.

Figure 4-2: Architecture of the Bounding Box Regressor Network. The final fully-connected layers in the original network are modified to introduce an additional head trained with Euclidean loss.



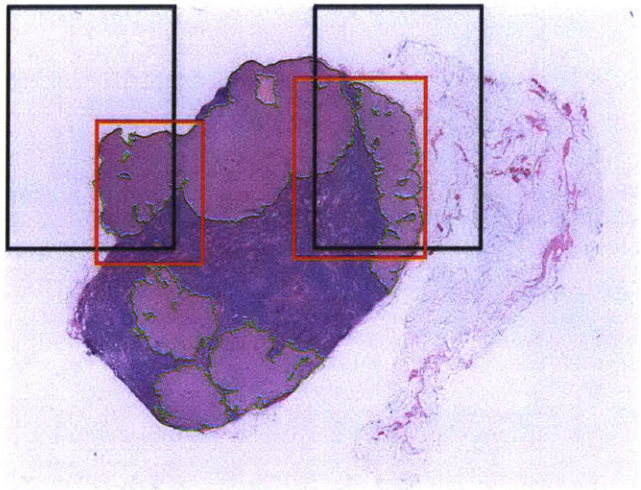
To generate object bounding box predictions, we simultaneously run the classifier and regressor networks across all locations and scales. Since these share the same feature extraction layers, only the final regression layers need to be recomputed after computing the classification network. The output of the final softmax layer for a tumor at each location provides a score of confidence that a tumor is present (though not necessarily fully contained) in the corresponding field of view. Thus we can assign a confidence to each bounding box. Training the regressors in a multi-scale manner is important for the across-scale prediction combination. Training on a single scale will perform well on that scale and still perform reasonably on other scales.

However training multi-scale will make predictions match correctly across scales and exponentially increase the confidence of the merged predictions.

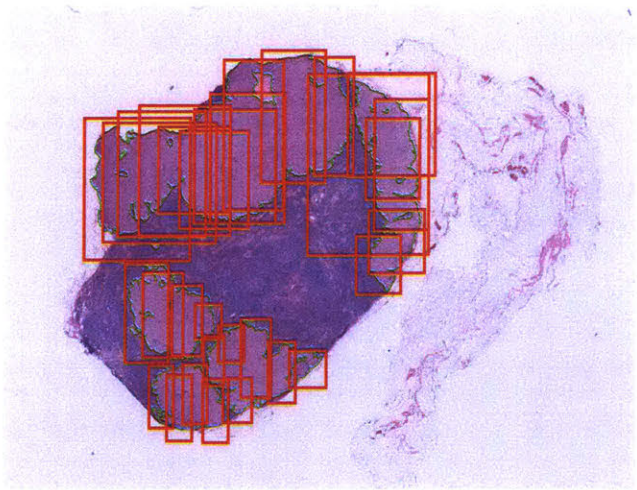
We combine the individual predictions via a greedy merge strategy applied to the regressor bounding boxes first introduced in [54].

1. Let C_s be the set of classes in the top k for each scale s, found by taking the maximum detection class outputs across spatial locations for that scale.
2. Let B_s be the set of bounding boxes predicted by the regressor network for each class in C_s, across all spatial locations at scale s.
3. Let $B = \cup_s B_s$
4. Repeat until convergence:
   - $(b_1^*, b_2^*) = argmin_{b_1 \neq b_2 \in B} MatchScore(b_1, b_2)$
   - Stop, if $MatchScore(b_1^*, b_2^*) > t$.
   - Else, let $B = B \backslash \{b_1^*, b_2^*\} \cup BoxMerge(b_1^*, b_2^*)$

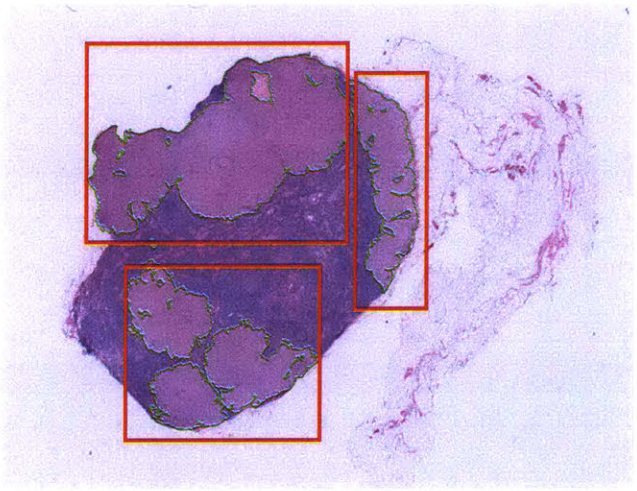
where $MatchScore$ computes the sum of the distance between centers of the two bounding boxes and the intersection area of the boxes and $BoxMerge$ compute the average of the bounding boxes coordinates. The final prediction is given by taking the merged bounding boxes with maximum class scores, which is computed by cumulatively adding the detection class outputs associated with the input windows from which each bounding box was predicted. An example illustrating the pipeline is shown in Figure 4-3



(a) Black rectangles mark the area under inspection at this scale, while red rectangles mark the predicted bounding box for that rectangle.



(b) High-confidence bounding boxes at multiple scale cluster around the tumors when visualized



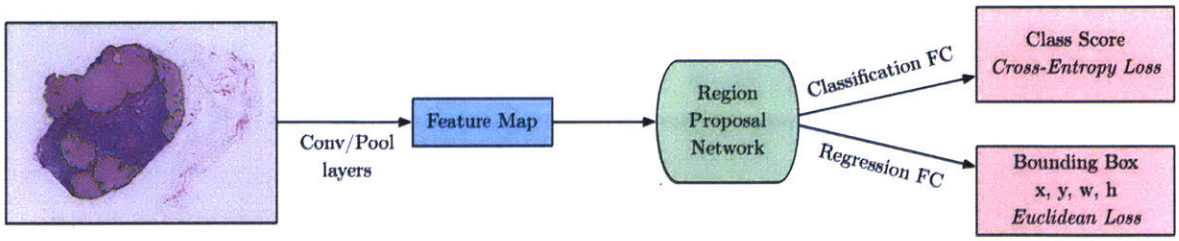
(c) Greedy merging of the boxes by confidence results in localization boxes for a particular scale.

Figure 4-3: The three stages of the localization-as-regression algorithm.

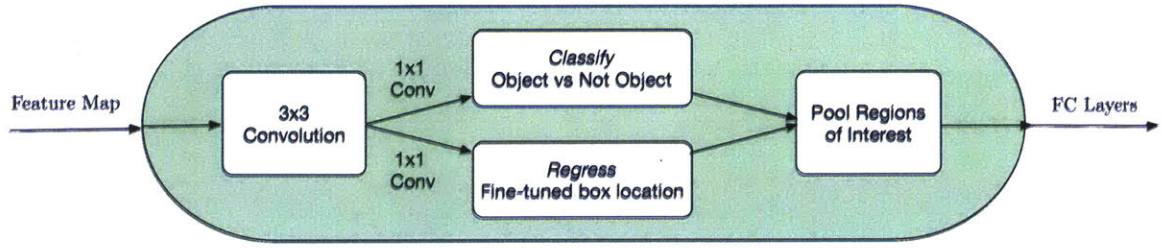
4.4 Region Proposal Network

To further examine joint learning methods for the localization task, we evaluate a region proposal network (first introduced in [50]) to hypothesize object locations. A RPN is a fully convolutional network that simultaneously predicts object bounds and objectness scores at each position. Using the so-called "attention mechanism", the RPN component tells the unified network where to look.

The implementation is divided into two modules (Figure 4-4). The first module is a deep fully convolutional network that proposes regions and the second module is the detector described earlier in Section 4.3, which uses the proposed regions. To generate region proposals, we slide a small network over the convolutional feature map output by the last shared convolutional layer. This small network takes as input a 3×3 spatial window of the input convolutional feature map. Each sliding window is mapped to a lower-dimensional feature, which is fed into two sibling fully connected layers - a box-regression layer and a box-classification layer.



(a) Modified bounding box regressor scheme with introduction of RPN.



(b) .

Figure 4-4: Schematic of the Fully Convolution structure of the RPN.

This approach is translation-invariant, and can easily be extended to the multi-scale setting with a pyramid of anchors [50]. However, we didn't observe significant boost in the performance by incorporating more than one scale - most of the discriminatory information for the pathology dataset appears to be at the highest magnification detail.

4.5 Results and Discussion

We compare the performance (Figure 4-5) of the three methods described above, along with a baseline of a very simple scheme of non-maxima supression which performs the following steps until convergence -

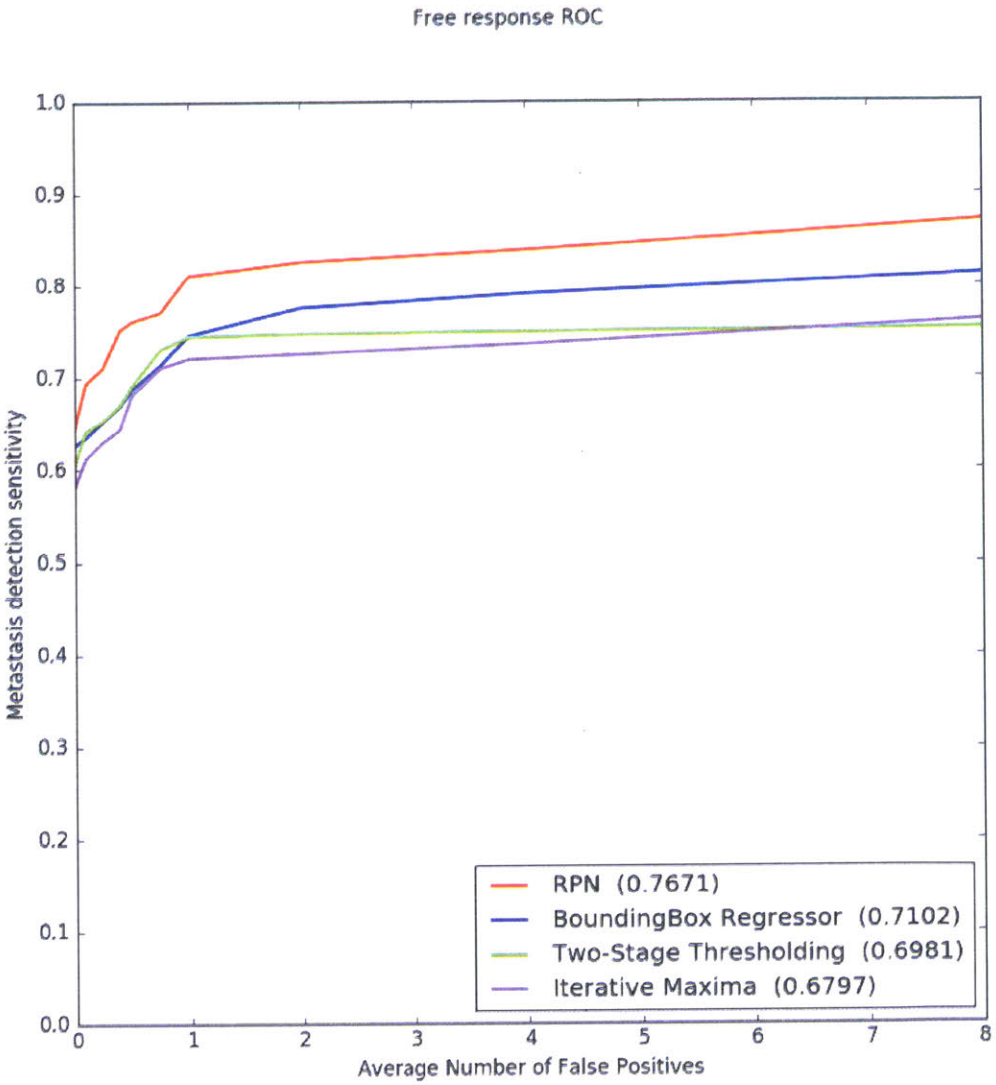
1. Take the maximum in the likelihood map as the probability of tumor in the entire image.
2. Set all values with a radius r of the maximum to 0.

The FROC metric is challenging because reporting multiple points per false positive region can significantly erode the score. We focused on the FROC as opposed to the AUC because there are approximately twice as many tumors as slides, which improves the reliability of the evaluation metric.

We also experimented with models pre-trained on ImageNet [52] which improved convergence speed in most cases, but failed to improve FROC, possibly due to the large difference in the kind of information present in the natural scenes of ImageNet vs the domain-specific pathology images, limiting the transferability.

Ensembling methods (averaging predictions from different augmentations, and averaging predictions of multiple independently-trained models) failed to provide a significant increase in the FROC score.

Figure 4-5: Performance Characteristics on the Lesion Localization Task



Chapter 5

Algorithms for Patient Stage Determination

Moving from the detection and localization tasks described previously to patient-level analysis requires combining the detection and classification of metastases in multiple lymph node slides into one outcome: a pN-stage. This task has high clinical relevance and would normally require extensive microscopic assessment by pathologists of tens of slides per patient. An automated solution for assessing the pN-stage in breast cancer patients, would hold great promise to reduce the workload of pathologists, while at the same time, reduce the subjectivity in diagnosis.

5.1 Dataset

In addition to the whole-slide images of hematoxylin-and-eosin stained lymph node sections with lesion-level annotations, the dataset for patient-level analysis consisted of 5 lymph node slides each for 100 patients in the training set, and 5 lymph node

slides each for 100 patients in the test set, for a total of 1000 slides. Each slide had a corresponding severity stage (negative, micro-metastasis, macro-metastasis, isolated tumor cells (ITC)), but the ITC were not annotated exhaustively. Each patient had a corresponding pN stage label. This data was collected from 5 different laboratories by the providers and all ground-truth annotations were carefully prepared by expert-level pathologists [22].

In actual clinical practice, tens of lymph nodes are surgically removed and processed in the pathology laboratory. However, the size of the dataset would grow beyond reasonable size if all the slides were shared, hence the providers forged artificial patients, with 5 slides provided for each patient and each slide corresponding to exactly one node.

5.2 Evaluation Metrics

Internationally, the TNM system is widely accepted [57] as a means to classify the extent of cancer spread in patients with a solid tumour. It is one of the most important tools for clinicians to help them select the suitable treatment option and to obtain an indication of prognosis [9]. This prognosis usually worsens as the cancer spreads [53]. In breast cancer, TNM staging takes into account the size of the tumour (T), whether the cancer has spread to the regional lymph nodes (N), and whether the tumour has metastasised to other parts of the body (M). The dataset we had access to [22] focused on the pathologic N-stage (pN-stage) of the patients, which could be one of five classes.

Thus, the evaluation metric used was a five-class quadratic weighted Cohen's kappa coefficient, which measures inter-rater agreement with the pathologist provided ground truth. [11]

$$\kappa = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}x_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}m_{ij}}$$

where k is the number of classes (pN-stages) and w_{ij}, x_{ij}, and m_{ij} are elements in the weight, observed, and expected matrices, respectively. The weights in the quadratic set are defined as:

$$w_i = 1 - \frac{i^2}{(k-1)^2}$$

The score is zero if the procedure does no better than random chance.

5.3 Rule-based approach

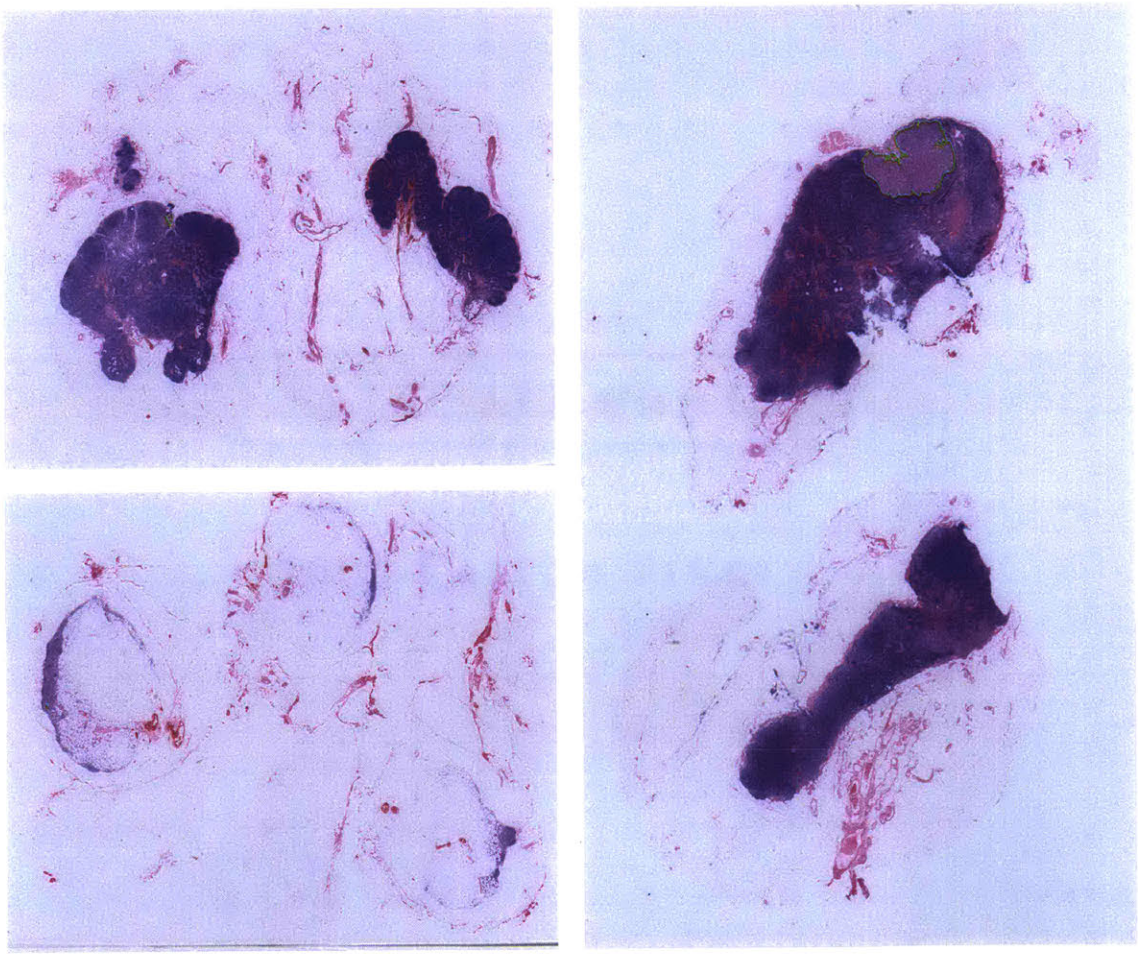
To compose a pN-stage, pathologists counts the number of positive lymph nodes (i.e. nodes with a metastasis). There are four categories a lymph node slide can be classified into:

Slide category	Rule
Macro-metastases	Metastases greater than 2.0 mm.
Micro-metastases	Metastases greater than 0.2 mm or more than 200 cells, but smaller than 2.0 mm.
Isolated Tumor Cells (ITC)	Single tumor cells or a cluster of tumor cells smaller than 0.2 mm or less than 200 cells.
Negative	None of the above.

Table 5.1: Metastasis classification rules

ITC is strictly not a metastasis and therefore lymph nodes containing only ITC are not counted as positive lymph nodes. However, pathologists are required to report on ITC when no macro-metastases or micro-metastases were found in a patient's lymph nodes.

Figure 5-1: Examples of lymph node sections with different metastasis classifications (micro-metastasis top-left, macro-metastasis right, isolated tumor cells bottom-left). The extremely small size of micro-metastases and ITC on this zoom level illustrate the tediousness a pathologist faces while examining each small subregion at a higher magnification.



s

Pathologic lymph node classification (pN-stage) is then defined as in 5.2.

Patient category	Rule
pN0	No micro-metastases or macro-metastases or ITCs found.
pN0(i+)	Only ITCs found.
pN1mi	Micro-metastases found, but no macro-metastases found.
pN1	Metastases found in 1-3 lymph nodes, and at least one is a macro-metastasis.
pN2	Metastases found in 4-9 lymph nodes, and at least one is a macro-metastasis.

Table 5.2: pN-stage classification system rules

This methodology is a simplified version of the official recommendations of the Union for International Cancer Control and the American Joint Committee on Cancer [46], due to lack of fine-grained data on lymph node's location of origin and molecular techniques like RT-PCR.

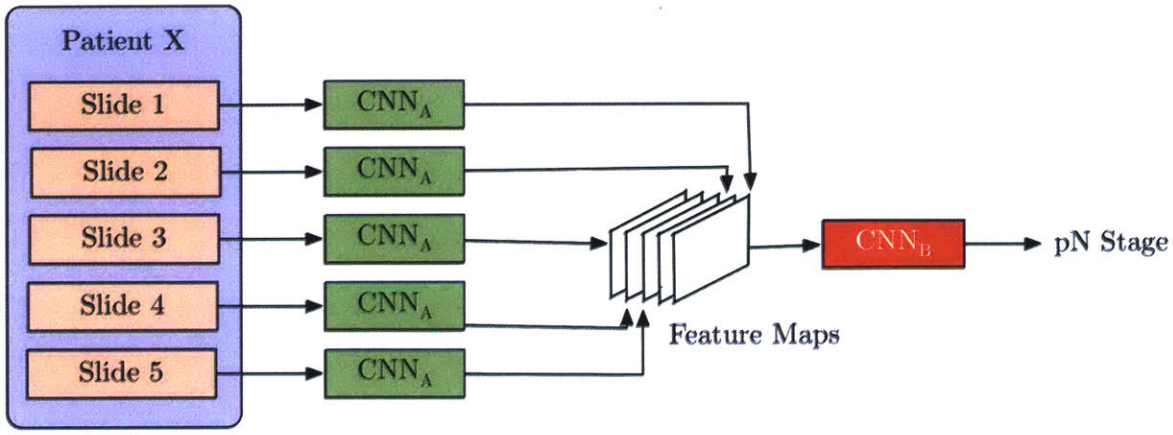
We modify our post-processing step defined in Section 3.3.4 to classify into 4 categories - macrometasis, micrometastasis, ITC and negative, instead of just the two as earlier. We retrain the random forest classifer and then apply the rules as described above to get a final pN stage for the patient.

5.4 Joint Learning of slide label and pN stage

The pN stages provided in the dataset are consistent with the individual whole-slide labels, and the rules for determining the stage as described earlier. The rule-based approach's limitations are only (at least in this context) caused by imperfect classification into the four classes.

To explore the possibility of a model which bypasses the explicit step of learning the individual whole-slide label, and learns the pN stage directly from the individ-

Figure 5-2: Architecture for directly learning the pN stage of the patient from whole-slide images, rather than first explicitly learning the slide class label.



ual slides, we restructure our network architecture (Figure 5-2) We run the patch-classification and recombination into heatmaps as before, but now train a new deep network (ResNet-34) with each patient's 5 slides as a stacked input image of 5 channels, to classify between the five pN stages. We boost the weight decay to avoid overfitting with the low number of examples.

5.5 Results and Discussion

We obtained the scores of Table 5.3 in the Cohen's Kappa evaluation metric. We note that the joint-learning approach boosted the score significantly, indicating that directly using the pN stage as the variable to be corrected reduced the build-up of errors from explicitly making an intermediate classification.

In particular, we also noted that the isolated tumor cells (ITC) were not exhaustively annotated in the ground truth set. If we replaced the labels of the slides marked as ITC to negative or to micro-metastases, we observe a minor improvement in performance.

The Kappa metric is challenging since even with quadratic weighing, the proximity

Method	κ score	κ score with ITC as negative	κ score with ITC as microm.
Joint-learning	0.8758	0.8921	0.8914
Rule-based approach	0.8386	0.8452	0.8435
Average Pathologist [17]	0.7 (Estimate)	-	-
Random Chance	0.0000	0.0000	0.0000

Table 5.3: Performance on the patient stage determination task

between individual classes is unaccounted for. Moreover, the highly critical nature of detecting all positives, at some cost of false positives, can only be captured by adjusting the weights accordingly. Nevertheless, the performance of the algorithms is a good indicator of the inter-rater agreement, which surpasses the average pathologist's score of 0.7 [17].

Chapter 6

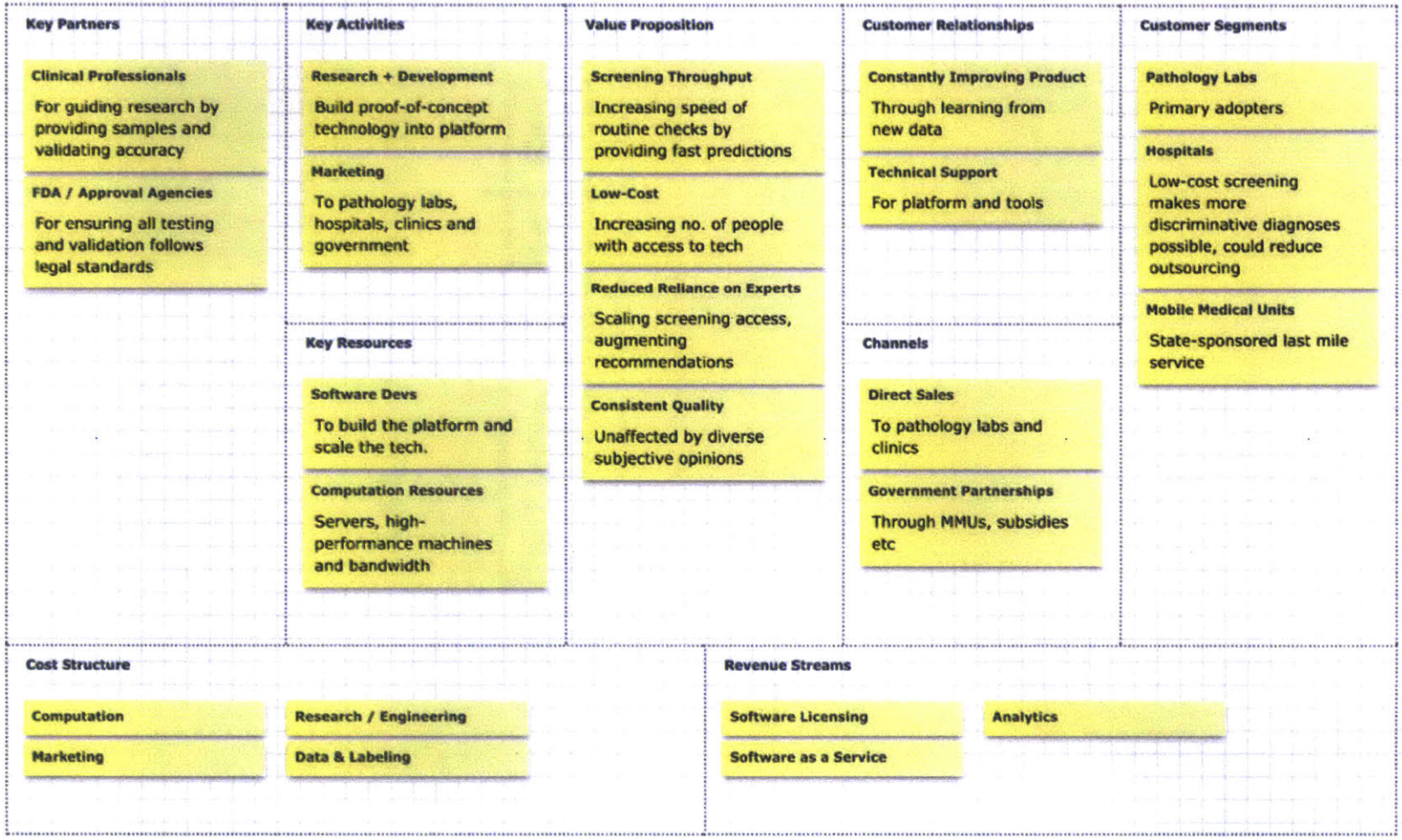
Analysis of Deployment Strategies

An automated histopathological analysis platform could help the lives of millions of Indians who require screening and diagnostic services every year, by enabling a reduction in reliance on hard-to-acquire experts, increase in throughput, and improvement in efficiency. It could be effectively deployed into the diagnostic pipeline of existing organized pathology lab chains, recuperating initial costs effectively over time while simultaneously getting better in terms of accuracy as access to more continuous data is made available. We present here a quantified articulation of a business model including activities, resources, customers, channels, costs, revenue streams and the value added.

6.1 Quantified Value Articulation

The healthcare industry amounts to over $80 billion in India, with over $1 billion directly coming from the domestic pathology market (growing at 20% compounded annual growth rate), mostly in Tier I cities [19]. This growth is heavily driven by

Figure 6-1: A business model canvas for deploying and commercializing the technologies and algorithms described in the previous chapters



changing disease profile of the population and consequent interest in screening and follow-up, including cancer markers in the recent past [37]. It is estimated to account for 2.5% of the healthcare delivery market [33]; conversely, it is also estimated to comprise 52% of the medical device market [37]. Like other healthcare services, laboratory testing is unevenly distributed across India. Research suggests that only one-third of the rural population has access to diagnostic centers and perhaps only one-half of public-sector facilities offer treatment for chronic illnesses [18]. The costs of delayed, inaccurate or inconsistent pathological screening and diagnostics invariably seep into the larger figure.

There are several problems to be confronted in this sector –

1. Only about 1% of the laboratories are accredited (e.g. by the National Accreditation Board of Laboratories) or certified (e.g. by the College of American Pathologists). There is no quality regulation in this sector; center operators do not require registration or certification of lab testing instruments or clinical processes.

2. There is little use of IT, automated equipment, and quality or statistical analyses run on this equipment outside of the organized sector.

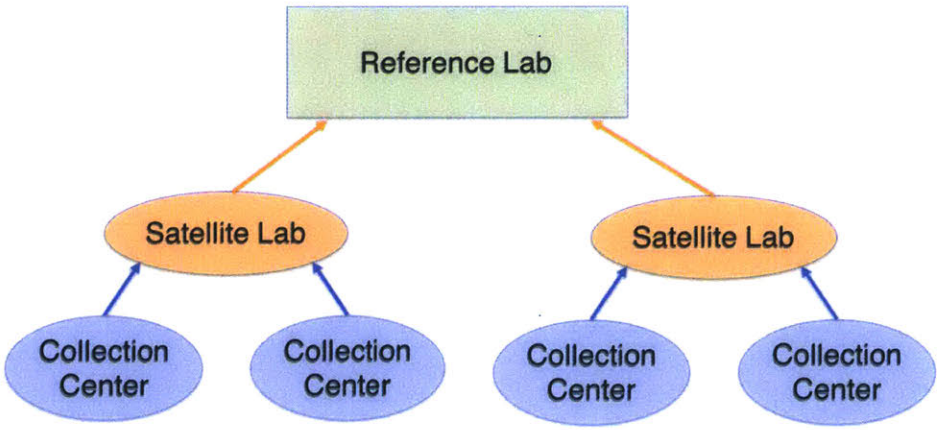
3. As a result of the first two factors, there is reportedly a large variation in laboratory performance across India, which has significant negative consequences for patient diagnosis, treatment and follow-up.

A pathology expert in India earns around $30,000-$80,000 per year, far above average of comparable fields of expertise. This directly stems from low supply and high demand, with each pathology test and analysis costing anywhere between $100-$300, where over 75% of the cost is due to the analyses rather than the physical collection and processing. If an automated system is used as a first line of defense, and assuming

75% of the cases can be classified as benign or malignant by the system, it can drive costs down to $50-$150 if amortized over the cost of setting up the system. This can translate into enough capability to hire over 2 to 3 times more experts, helping pathology labs scale up their operations and expand into newer markets.

Currently, to analyze tissue samples of patients for indications of conditions like cancer, anemia, tuberculosis etc., they are usually extracted at a local hospital or clinic, and then either physically sent to a pathological expert without imaging (1 week), or digitized with a microscope at a collection center and sent to a pathological expert online. (usually 2-3 days). An automated analysis system, if combined with the collection / digitization facilities can reduce the turnaround time to a few minutes for most cases, thereby heavily increasing throughput and helping direct care to those who need it most urgently.

Figure 6-2: The collection center model for histopathological analyses in India.



The quality of analysis and recommendations varies wildly with demography, expertise and cost / time currently, but having an automated system can help inform consistent metrics for counsel, introducing equity in analyses regardless of the patient's conditions or financial status.

6.1.1 Channels

Direct sales to pathology labs and clinics could be the biggest channel for deployment of the service. Each lab spends over $15-$40 per hour on an expert and a licensing model which make the software available as a service for remote use can be priced at a fraction, say $5-$10 per hour, and still be profitable because of the mostly one-time (initial) nature of the investment, which can be amortized over a long period and the costs recuperated.

An alternative to direct sales could be through partnerships, especially with the central and state governments, which are heavily invested in providing improvements to health care, increase life expectancy and provide connectivity to hard-to-reach and thus untapped geographies. Subsidies to the order of 30-40% aren't unheard of and can help drive adoption by bigger commercial players.

6.1.2 Revenue Streams

The primary revenue streams would be based on providing the platform software, either on a licensing model or as a remote service, with the pricing based on number of analyses performed or amount of data analyzed. A price-point of $10 per patient analysis could result in over $10 million of revenue per year, while still recovering the initial cost of investment, and yet being cost-effective for the pathology labs. Also, as the accuracy of the platform improves with more available data to analyze, its value would appreciate, rather than depreciate, over time.

A de-identified demographic analytics layer, on top of the actual assessment platform could also be immensely valuable as a tool for policy makers, hospital administrators, and consultants. This could potentially form an alternative secondary revenue stream.

6.1.3 Cost Structure

The computation resources required to train statistical models and host the cloud-platform would be a key investment initially, for research and development. A powerful machine for the former could cost $10,000 - $15,000 and the latter can cost $1,000 - $2,000. The number of servers would need to scale to handle the load of servicing once the platform is deployed and could handle 3-5 active analyses simultaneously, resulting in $10,000 - $15,000 on machines per pathology lab, if 25 technicians utilize it at once.

The research and engineering efforts would also require a significant investment in terms of expertise in machine learning systems and scalable cloud-based software development. A team of 3-5 competent engineers, could require an investment of $300,000 - $500,000 per year in the development phases. The sales and marketing end of things would also require effective personnel for business development and support, and depending on the number of clients / customers, could cost over $500,000 per year.

Access to datasets and corresponding annotations (to build statistical models from) is probably the most uncertain investment in terms of pricing at this point. There are a few public datasets available for certain histopathological tasks but a commercial solution would inevitably need access to a sizable amount of custom annotated data. There are a few companies (e.g. Advanced Tissue Services, Asuragen etc.) which provide samples, digitized or otherwise, at prices ranging from $0.20-$2 per slide. An analysis of the tradeoff between amount of data required vis a vis expected performance would require more experiments at this point.

The bulk of the initial effort would be in research and development for the platform, including but not limited to data collection, machine learning engineering, web server development and cloud-based scaling. Investment for the same could amount to $1

million - $3 million as computed earlier, but would be recovered over time from sales.

The subsequent investment would be in marketing the service to pathology labs and clinics, as either a SaaS or as a software licensing model. The costs involved with business development are highly variable but are expected to be 30-60% of the engineering costs, based on historical data of similar software-oriented solutions.

6.2 Customer Segments

The key customer segments include organized and unorganized diagnostic labs, hospitals and clinics, and state-sponsored initiatives like mobile medical units and subsidies.

	Organized Labs	**Unorganized Labs**	**Hospitals**	**Mobile Medical Units**
End User	Pathologists & Technicians	Pathologists & Technicians	Clinicians & Nurses	State Government Employees
Buyer	Lab Chain	Lab Owner	Hospitals	State Government
Applications	Screening & Diagnostics	Screening & Diagnostics	Preliminary Screening & Diagnostics	Last-mile health care connectivity
Benefits	Penetration into Tier II / III markets	Competitive advantage over organized labs	Not reliant on path labs	Faster turnaround time
Market Size	$1b total market, rapid expansion	80% of lab market	Hospital services $2b	1500+ units, 38m+ patients / year

Table 6.1: Breakdown of key customer segments

The organized sector has arisen in recent years to build a national network and develop a more capital-intensive segment ducting high-margin testing. To develop

this network, firms have begun to enter Tier II and Tier III cities, typically via acquisitions and brownfield projects. It is by far the most consolidated segment, with clear key players, and a predictable growth trajectory. They have employed variants of the hub-and-spoke model in which they operate central reference labs in large metropolitan areas and partner with satellite and feeder labs as well as hospital collection centers.

Automated solutions and systems will only benefit them by reducing inefficiencies, providing a consistent quality of service and keeping costs low, and thus they might be more willing to help validate and test experimental technology. Each lab employs 1000-3000 experts, and the big chains (5-8 in number) have tens to hundreds of reference labs spread around the country, but mostly concentrated in major cities. Some of these labs include –

Pathology Lab	Number of Labs	Collection Centers
Dr Lal Path Labs	35	600
Metropolis	50	350
SRL Diagnostics	35	800
Piramal Diagnostics	104	300

Table 6.2: Characteristics of notable diagnostic lab chains

The majority of the market share however (estimated to be as high as 85-90%) is captured by the unorganized sector [56]. The small operations have proliferated due to lack of regulation and requirements for accreditation in this sector; they also have comparably lower costs and referral ties to physicians, who may receive kickbacks for patient testing, which may explain their prevalence [37]. Incorporating automated analyses systems can help them compete better with bigger players who might have more access to expert talent.

The laboratory plays a critical role in the delivery of Indian healthcare; one estimate suggests that as much as 70% of all medical decisions are based on laboratory re-

sults. Hospitals, particularly those with fewer than 250 beds (which comprise the vast majority of Indias facilities), often outsource much of their diagnostic testing to large freestanding laboratories. This outsourcing approach, sometimes referred to as hospital laboratory management (HLM), seeks to take advantage of aggregated test volumes and scale economies. Only about 10% of Indian hospitals (primarily private facilities) have their own laboratories. However, if an automated preliminary screening system is deployed, they would be able to reduce the reliance on outsourcing for all cases, and only consult pathology labs for low-confidence or high-risk cases.

There has been a recent aggressive push towards deploying mobile medical vans / units (MMUs) at different levels by state governments. About 1500 of them are currently active across the country, primarily in Orissa, Jharkhand, Maharashtra and a few other states, with rapid expansion plans [62]. Each of them services an average of 25,000 patients per year providing last mile access. Theyre outfitted with X-Ray machines, ultrasound, blood collection and tissue collection facilities. The service costs an amortized Rs. 150 per patient. They suffer from sample degradation during transportation and delays in obtaining results. They would benefit heavily if faster analyses of collected samples was available.

Each segment has their pros and cons in terms of initial deployment support and offerings, but the organized diagnostic lab sector could be a good starting point since they have a number of benefits to reap from such a system and market forces help in incentivizing it, thus positioning the project in a favorable position to move forward.

Chapter 7

Conclusions

7.1 Summary

The histological assessment of human tissues has emerged as a key challenge for detection and treatment of multiple clinical conditions, including cancer. Experts analyze slide images in order to characterize samples by performing tasks like whole-slide classification, tumor localization etc. whose results inform the diagnosis. But a mismatch between demand and supply of experts, especially in the Indian setting and the rural context, has led to suboptimal quality and turnaround times.

The proposed automated platform for large-scale histopathological screening presented in this thesis can help automate preliminary analysis tasks, massively reducing reliance on expensive experts, increasing throughput and accuracy, providing consistent results over widely varying cases, while still allowing humans to verify and have the final word on any grading. It can help reduce inefficiencies in scaling and expanding diagnostic services, while still being compatible with existing diagnostic pipelines and procedures. It could be effectively deployed into the diagnostic pipeline of ex-

isting organized pathology lab chains, recuperating initial costs effectively over time while simultaneously getting better in terms of accuracy as access to more continuous data is made available.

In Chapter 3, we described algorithms to classify whole-slide images of sentinel lymph node tissues extracted from patients into normal or tumorous. We described a classification pipeline, with pre-processing of background removal, stain normalization and data augmentation, followed by patch-wise classification with a deep-neural network to output a likelihood, and finally post-processing of the likelihood map with a random forest trained on the geometric properties of the tumor to give a class label.

In Chapter 4, we detailed algorithms to localize the lesion in the whole-slide image, using three approaches - a two-stage thresholding model which oversamples tumor-adjacent regions to achieve high discriminability, a bounding box regressor model which learns to predict the bounding box at multiple scales which are then merged greedily weighted by confidence, and a region proposal network addition to the regressor which identifies regions of high interest.

In Chapter 5, we discussed algorithms for determining the clinical stage of the patient. We contrasted the traditional rule-based approach with a novel joint-learning method which bypasses the intermediate explicit labeling of individual slides in favor of an end-to-end pN stage classification.

Finally, in Chapter 6, we examined and articulated deployment strategies for a healthcare software solution encompassing these algorithms in the Indian setting, including the market state, distribution channels, revenue streams, associated cost structure and key customer segments.

7.2 Future Work

The algorithms described in this thesis perform well for the tasks for which they are trained, but lack they currently lack the breadth of knowledge and experience of human pathologists for example, being able to detect other abnormalities that the model has not been explicitly trained to classify (e.g. inflammatory process, autoimmune disease, or other types of cancer). Gathering more data to that end and tuning the models to recognize more pathological cases will improve performance on such 'hitherto unspecified' tasks and make for a more robust real-world solution.

Bibliography

[1] Erwin H Ackerknecht et al. Rudolf virchow: Doctor, statesman, anthropologist. *Rudolf Virchow: Doctor, Statesman, Anthropologist.*, 1953.

[2] Max K Agoston. *Computer graphics and geometric modeling*, volume 1. Springer, 2005.

[3] Sophia K Apple. Sentinel lymph node in breast cancer: review article from a pathologists point of view. *Journal of pathology and translational medicine*, 50(2):83, 2016.

[4] Narayanaswamy Balakrishnan. *Handbook of the logistic distribution.* CRC Press, 2013.

[5] Andrew H Beck, Ankur R Sangoi, Samuel Leung, Robert J Marinelli, Torsten O Nielsen, Marc J van de Vijver, Robert B West, Matt van de Rijn, and Daphne Koller. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Science translational medicine*, 3(108):108ra113–108ra113, 2011.

[6] Babak Ehteshami Bejnordi, Geert Litjens, Nadya Timofeeva, Irene Otte-Höller, André Homeyer, Nico Karssemeijer, and Jeroen AWM van der Laak. Stain specific standardization of whole-slide histopathological images. *IEEE transactions on medical imaging*, 35(2):404–415, 2016.

[7] Olivier Bousquet and Léon Bottou. The tradeoffs of large scale learning. In *Advances in neural information processing systems*, pages 161–168, 2008.

[8] Fadi Brimo, Luciana Schultz, and Jonathan I Epstein. The value of mandatory second opinion pathology review of prostate needle biopsy interpretation before radical prostatectomy. *The Journal of urology*, 184(1):126–130, 2010.

[9] Steven L Chen, Francesca M Hoehne, and Armando E Giuliano. The prognostic significance of micrometastases in breast cancer: a seer population-based analysis. *Annals of surgical oncology*, 14(12):3378–3384, 2007.

[10] Dan C Cireşan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Mitosis detection in breast cancer histology images with deep neural networks. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 411–418. Springer, 2013.

[11] Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

[12] Ronan Collobert, Koray Kavukcuoglu, and Clément Farabet. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*, number EPFL-CONF-192376, 2011.

[13] Angel Cruz-Roa, Ajay Basavanhally, Fabio González, Hannah Gilmore, Michael Feldman, Shridar Ganesan, Natalie Shih, John Tomaszewski, and Anant Madabhushi. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. In *SPIE medical imaging*, pages 904103–904103. International Society for Optics and Photonics, 2014.

[14] Brian J Czerniecki, Alice M Scheff, Linda S Callans, Francis R Spitz, Isabelle Bedrosian, Emily F Conant, Susan G Orel, Jesse Berlin, Cynthia Helsabeck, Douglas L Fraker, et al. Immunohistochemistry with pancytokeratins improves the sensitivity of sentinel lymph node biopsy in patients with breast carcinoma. *Cancer*, 85(5):1098–1103, 1999.

[15] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[16] Stephen B Edge and Carolyn C Compton. The american joint committee on cancer: the 7th edition of the ajcc cancer staging manual and the future of tnm. *Annals of surgical oncology*, 17(6):1471–1474, 2010.

[17] Joann G Elmore, Gary M Longton, Patricia A Carney, Berta M Geller, Tracy Onega, Anna NA Tosteson, Heidi D Nelson, Margaret S Pepe, Kimberly H Allison, Stuart J Schnitt, et al. Diagnostic concordance among pathologists interpreting breast biopsy specimens. *Jama*, 313(11):1122–1132, 2015.

[18] Nora Engel, Gayatri Ganesh, Mamata Patil, Vijayashree Yellappa, Nitika Pant Pai, Caroline Vadnais, and Madhukar Pai. Barriers to point-of-care testing in india: results from qualitative research across different settings, users and major diseases. *PLoS One*, 10(8):e0135112, 2015.

[19] Ernst and Young. India healthcare. india brand equity foundation, 2017.

[20] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[21] National Breast Cancer Foundation. http://www.nationalbreastcancer.org/metastatic-breast-cancer, 2017.

[22] Oscar Geessink, Péter Bándi, Geert Litjens, and Jeroen van der Laak. Camelyon17: Grand challenge on cancer metastasis detection and classification in lymph nodes, 2017.

[23] Farzad Ghaznavi, Andrew Evans, Anant Madabhushi, and Michael Feldman. Digital imaging in pathology: whole-slide imaging and beyond. *Annual Review of Pathology: Mechanisms of Disease*, 8:331–359, 2013.

[24] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.

[25] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[26] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22):2402–2410, 2016.

[27] Metin N Gurcan, Laura E Boucheron, Ali Can, Anant Madabhushi, Nasir M Rajpoot, and Bulent Yener. Histopathological image analysis: A review. *IEEE reviews in biomedical engineering*, 2:147–171, 2009.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[30] Robert Hecht-Nielsen et al. Theory of the backpropagation neural network. *Neural Networks*, 1(Supplement-1):445–448, 1988.

[31] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[32] Humayun Irshad, Antoine Veillard, Ludovic Roux, and Daniel Racoceanu. Methods for nuclei detection, segmentation, and classification in digital histopathology: a reviewcurrent status and future potential. *IEEE reviews in biomedical engineering*, 7:97–114, 2014.

[33] India Semiconductor Association (ISA). Current status and potential for medical electronics in india., 2010.

[34] Shabnam Jaffer and Ira J Bleiweiss. Evolution of sentinel lymph node biopsy in breast cancer, in and out of vogue? *Advances in anatomic pathology*, 21(6):433–442, 2014.

[35] Andrew Janowczyk and Anant Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics*, 7, 2016.

[36] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[37] KPMG. Excellence in diagnostic care, 2011.

[38] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[39] Vinay Kumar, Abul K Abbas, Nelson Fausto, and Jon C Aster. *Robbins and Cotran pathologic basis of disease*. Elsevier Health Sciences, 2014.

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

[41] Geert Litjens, Clara I Sánchez, Nadya Timofeeva, Meyke Hermsen, Iris Nagtegaal, Iringo Kovacs, Christina Hulsbergen-Van De Kaa, Peter Bult, Bram Van Ginneken, and Jeroen Van Der Laak. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific reports*, 6, 2016.

[42] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[43] Gary H Lyman, Armando E Giuliano, Mark R Somerfield, Al B Benson III, Diane C Bodurka, Harold J Burstein, Alistair J Cochran, Hiram S Cody III, Stephen B Edge, Sharon Galper, et al. American society of clinical oncology guideline recommendations for sentinel lymph node biopsy in early-stage breast cancer. *Journal of clinical oncology*, 23(30):7703–7720, 2005.

[44] Gary H Lyman, Sarah Temin, Stephen B Edge, Lisa A Newman, Roderick R Turner, Donald L Weaver, Al B Benson III, Linda D Bosserman, Harold J Burstein, Hiram Cody III, et al. Sentinel lymph node biopsy for patients with early-stage breast cancer: American society of clinical oncology clinical practice guideline update. *Journal of Clinical Oncology*, 32(13):1365–1383, 2014.

[45] Raouf E Nakhleh. Error reduction in surgical pathology. *Archives of pathology & laboratory medicine*, 130(5):630–632, 2006.

[46] American Joint Committee of Cancer. https://cancerstaging.org/references-tools/quickreferences/documents/breastmedium.pdf, 2017.

[47] OptraScan. http://www.optrascan.com/, 2017.

[48] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

[49] Stephen S Raab, Dana Marie Grzybicki, Janine E Janosky, Richard J Zarbo, Frederick A Meier, Chris Jensen, and Stanley J Geyer. Clinical impact and frequency of anatomic pathology errors in cancer diagnoses. *Cancer*, 104(10):2205–2213, 2005.

[50] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[51] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2564–2571. IEEE, 2011.

[52] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[53] Sepideh Saadatmand, Reini Bretveld, Sabine Siesling, and Madeleine MA Tilanus-Linthorst. Influence of tumour stage at breast cancer detection on survival in modern times: population based study in 173 797 patients. *bmj*, 351:h4901, 2015.

[54] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[55] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[56] Dhirendra Singh. The needle of diagnostics, 2011.

[57] Leslie H Sobin and Irvin D Fleming. Tnm classification of malignant tumors, (1997). *Cancer*, 80(9):1803–1804, 1997.

[58] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[59] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[60] Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: image processing in python. *PeerJ*, 2:e453, 2014.

[61] Mitko Veta, Paul J Van Diest, Stefan M Willems, Haibo Wang, Anant Madabhushi, Angel Cruz-Roa, Fabio Gonzalez, Anders BL Larsen, Jacob S Vestergaard, Anders B Dahl, et al. Assessment of algorithms for mitosis detection in breast cancer histopathology images. *Medical image analysis*, 20(1):237–248, 2015.

[62] Wharton. Mobile medical vans overcoming indias last mile health care challenges, 2013.

[63] Kun-Hsing Yu, Ce Zhang, Gerald J Berry, Russ B Altman, Christopher Ré, Daniel L Rubin, and Michael Snyder. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature Communications*, 7, 2016.