Providing Prognostic and Diagnostic Tools towards Melanoma Diagnosis

Shima Nofallah

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:
Linda Shapiro, Chair
Joann Elmore
Jenq-Neng Hwang

Program Authorized to Offer Degree:
Electrical and Computer Engineering

University of Washington

Abstract

Providing Prognostic and Diagnostic Tools towards Melanoma Diagnosis

Shima Nofallah

Chair of the Supervisory Committee:
Linda Shapiro
Paul G. Allen School of Computer Science & Engineering

The number of melanoma diagnoses has increased dramatically over the past three decades, outpacing almost all other cancers. Nearly 1 in 4 skin biopsies are of melanocytic lesions, highlighting the clinical and public health importance of correct diagnosis. Pathologists analyze biopsy material at both the cellular and structural level to determine diagnosis and cancer stage. Deep learning image analysis methods may improve and complement current diagnostic and prognostic capabilities. Mitotic figures are surrogate biomarkers of cellular proliferation that can provide prognostic information; thus, their precise detection is an important factor for clinical care. In addition, semantic segmentation of clinically important structures in skin biopsies is a crucial step toward an accurate diagnosis. We aim to provide prognostic and diagnostic information that consists of the detection of cellular level entities, clinically important structures, and other important factors in the diagnosis of skin biopsy images. This dissertation contains four main projects on melanocytic lesion biopsy images: mitotic figure classification, semantic segmentation of clinically important tissue structures, classification of segmented dermal nests, and improving whole slide image diagnosis using segmentation masks.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

So many wonderful people have helped me walk this far, and I am grateful for every one of them. First, I would like to express my appreciation for the mentorship of my advisor, Prof. Linda Shapiro. The opportunity you gave me forever changed my life. You are a woman of courage, independence, and ambition, and you are such an inspiration for me. I am honored to have had the chance to work with you.

Dr. Joann Elmore, you guided me through the field of cancer research and provided me with the tools I needed to become a better researcher. I deeply appreciate the support you kindly offered me through these years. To our amazing team of pathologists, Dr. Stevan Knezevich, Dr. Caitlin May, Dr. Oliver Chang, and Dr. Mojgan Mokhtari: my sincere appreciation to you all for patiently teaching me about pathology and offering your valuable time to answer my endless questions. I would also like to thank my collaborators Dr. Ezgi Mercan, Dr. Sachin Mehta, Wenjun Wu, Kechun Liu, Dr. Lisa Reisch, Dr. Donald Weaver, Dr. Daniela Witten, and Dr. Annie Lee for the opportunity of working with you and learning valuable lessons from each of you.

To my labmates who became my dear friends—Nicholas Nuechterlein, Dr. Beibin Li, Ananditha Raghunath, and Fatemeh Ghezloo—thank you for all the great conversations, kind words, and the support you offered.

I cannot imagine being where I am without having amazing friends who supported my growth in some of the most challenging seasons of my life: Ava, Elina, Negar, Sarah, Lucien, Abbas, Farnaz, Koosha, Saghar, Milad, and Ghazaleh, I appreciate you for being who you are. I would like to express my sincere gratitude to Maryam Namegh—thank you for encouraging me to stay true to my soul and for helping me figure out my strengths and weaknesses.

Finally, my highest gratitude goes to my family. My wonderful, selfless, amazing parents. I am in awe of your strength and your pure hearts. I could not have done any of this without your unconditional love and unwavering support. And to my beloved late brother, Pouya, your memory shines through my life every day. Thank you for the light, the passion, and the love you brought to my life. You are forever a piece of my heart.

DEDICATION

To my amazing parents, Abdollah and Zahra.

In memory of my dear brother, Pouya.

Chapter 1

INTRODUCTION

Cancer is a disease in which some of the body's cells grow uncontrollably and spread to other parts of the body. Normally, human cells grow and multiply to form new cells as the body needs them. Sometimes this orderly process breaks down, and abnormal or damaged cells grow and multiply when they shouldn't. These cells may form tumors, which are lumps of tissue. Tumors can be cancerous or non-cancerous. Cancer can start almost anywhere in the human body. There are more than 100 types of cancer, such as lung cancer, brain cancer, and skin cancer [56].

Skin cancer is the most common type of cancer. The main types of skin cancer are squamous cell carcinoma, basal cell carcinoma, and melanoma. Melanoma is much less common than the other types but much more likely to invade nearby tissue and spread to other parts of the body. Most deaths from skin cancer are caused by melanoma. Melanoma is a cancer that begins in melanocytes, which are specialized cells that make melanin (the pigment that gives skin its color). Most melanomas form on the skin, but melanomas can also form in other pigmented tissues, such as the eye [56]. The incidence of melanoma is rising faster than any other cancer [69, 36, 25].

The current gold standard for melanoma diagnosis is the microscopic examination of skin biopsies using hematoxylin and eosin (H&E) stained tissue sections; however, the histologic interpretation of melanocytic lesions is often inconsistent for pathologists. Pathologists' diagnostic interpretations are usually not verified by another pathologist. This is concerning, as diagnostic errors when interpreting melanocytic lesions occur much more frequently than in other tissues. Our research team has highlighted these challenges by demonstrating that pathologists disagree on up to 60% of cases of melanoma in-situ and T1a invasive melanoma,

which can lead to both overtreatment and undertreatment [19].

Whole slide digital imaging of pathology specimens can be used to create digitized slides, which in turn can be included in biorepositories or used in telepathology to enable diagnosis at a distance. By investigating the potential of computer technology to improve diagnoses using digitized slides and associated image characteristics, there is a possibility of providing clinical support tools for pathologists. Researchers have shown that automated diagnosis holds promise for improving accuracy and reproducibility in the diagnosis of histopathology [78, 21, 53, 92].

Layers of the skin on biopsy specimens are as follow: the epidermis on top, the dermis below it, and the hypodermis below that layer. Pathologists then look for two main types of evidence of invasive melanoma: cellular atypia and abnormal architecture. The presence, count, shape, size, and location of cellular entities such as nuclei, melanocytes, and mitotic figures play important diagnostic and prognostic roles in analyzing each skin biopsy image. Abnormality in structural levels such as dermis, epidermis, epidermal nests, and dermal nests also indicate the possibility of a cancerous component in the skin biopsy images.

Cellular level entities such as nuclei, melanocyte, and mitotic figures have diagnostic and prognostic value in the analysis of the skin biopsy images. The detection of nuclei is critical to the diagnosis of invasive melanoma. Melanocytic nuclei are the most important subset of the detected nuclei regions. Mitotic figures are surrogate biomarkers of cellular proliferation that can provide prognostic information; thus, their precise detection is an important factor for clinical care. Among the independent predictors of melanoma-specific survival, the mitotic rate is the strongest prognostic factor after tumor thickness [81].

Pathologists look at a skin biopsy slide and determine if its overall structure is normal, abnormal, or malignant. An automated approach for recognizing this structure requires a structure finder that looks for structural entities to determine if they are normal, abnormal (yet benign), or malignant. A normal (e.g., benign mole/nevus) structure contains well-spaced nests of melanocytes, sometimes forming a single layer with keratinocytes between them at the bottom of the epidermis. Abnormal structures are groups of melanocytes that can

be of non-uniform size or abnormally large, can form a bridge between two rete ridges (e.g. downward projections of epidermis into the dermis), can move up into the epidermis, or can appear in large groups in the dermis; such features are suggestive of malignancy.

There are variety of challenges in working with a medical dataset in general, especially the private medical sets. The private dataset is usually fairly small, which makes the process of training computational algorithms and validating those trained algorithms difficult. Another challenge in working with a medical dataset is the limitation of fine annotation as ground truth. The main reason is that annotation of large images is a labor-intensive task. This issue becomes especially pronounced in a medical image dataset in which expert annotation is the gold standard. In working on our skin biopsy dataset, both on the cellular level and structural level, we faced the challenge of scarce data annotation.

Numerous studies have introduced diagnosis models using histopathology images. In [52] utilizing variable-sized regions of interest, a CNN-based deep feature extraction framework was introduced to build slide-level feature representations via weighted aggregation of the patch representations. In another work, [38] utilized a self-supervised contrastive learning algorithm to extract representations from patches, and used an aggregator that models the relations of the instances in a dual-stream architecture with trainable distance measurement to train a MIL model called a Dual-stream Multiple Instance Learning Network (DSMIL). With the emergence of transformers in the filed of machine learning and computer vision, [9] proposed a Multiple Instance Learning (MIL) method that first selects the top-k patches, and then uses these patches for instance-learning and bag-representation learning. Clustering-constrained Attention Multiple (CLAM) instance learning is a deep-learning based weakly-supervised method introduced by [45] that uses attention-based learning to automatically identify sub-regions of high diagnostic value in order to accurately classify the whole slide, while also utilizing instance-level clustering.

There is various work related to the diagnosis of biopsy images of other types of cancers than melanoma, especially on breast histopathological Whole Slide Images (WSI) [51, 53, 57, 92]; however, related work for melanoma diagnosis using skin biopsy WSIs is very limited. There

are works on staining other than H&E, such as Ki-67 stain [1], which is not the standard stain for melanoma diagnosis in clinical care. Most existing work on melanoma diagnosis are either binary classification systems or are on not very challenging categories to distinguish [44, 93, 82, 29], which while still valuable, do not cover the whole spectrum of the diagnosis categories, especially the classes which are challenging for pathologists.

In this dissertation, four of our main projects on skin biopsy images are summarized: classification of mitotic figures, segmentation of important tissues in skin biopsy images, classification of dermal nests, and improvement of WSI diagnoses using segmentation masks. In Chapter 2, an overview of the main dataset that is used throughout this project is provided. In Chapter 3, the mitosis classification project is summarized. In Chapter 4, the segmentation of skin biopsy images using coarse and sparse annotation is described. In Chapter 5, a summary of classification of dermal nests is provided. In Chapter 6, the WSI diagnosis project and the potential of improving the diagnosis using segmentation masks from Chapter 4 and Chapter 5 are discussed. In chapter 7, the conclusions of all the projects and possible future work are discussed.

Chapter 2

M-PATH DATASET

Our dataset includes 240 hematoxylin and eosin (HE) stained slides of digitized skin biopsy images, acquired by a Bellevue, Washington dermatopathology laboratory for the M-Path study [19]. The study was approved by the Institutional Review Board at the University of Washington with protocol number STUDY00008506. This dataset contains melanocytic skin lesions from shave, punch, and excisional specimens. The cases can be classified into five different MPATH-Dx (Melanocytic Pathology Assessment Tool and Hierarchy for Diagnosis) simplified categories based on presumed risk of the lesion and suggested treatment recommendations [65]. Example diagnostic terms for each MPATH-Dx class are as follows: I) Mildly Dysplastic Nevi, II) Moderately Dysplastic Nevi, III) Melanoma in Situ and Severely Dysplastic Nevi, IV) Invasive Melanoma Stage T1a and V) Invasive Melanoma Stage T1b. Table 2.1 shows the distribution of the diagnostic categories of the M-path dataset. Figure 2.1 shows an example of three different WSIs in the M-Path dataset.

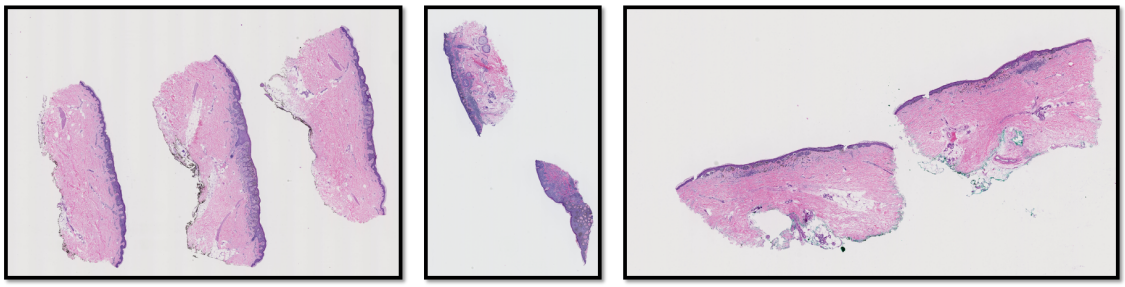


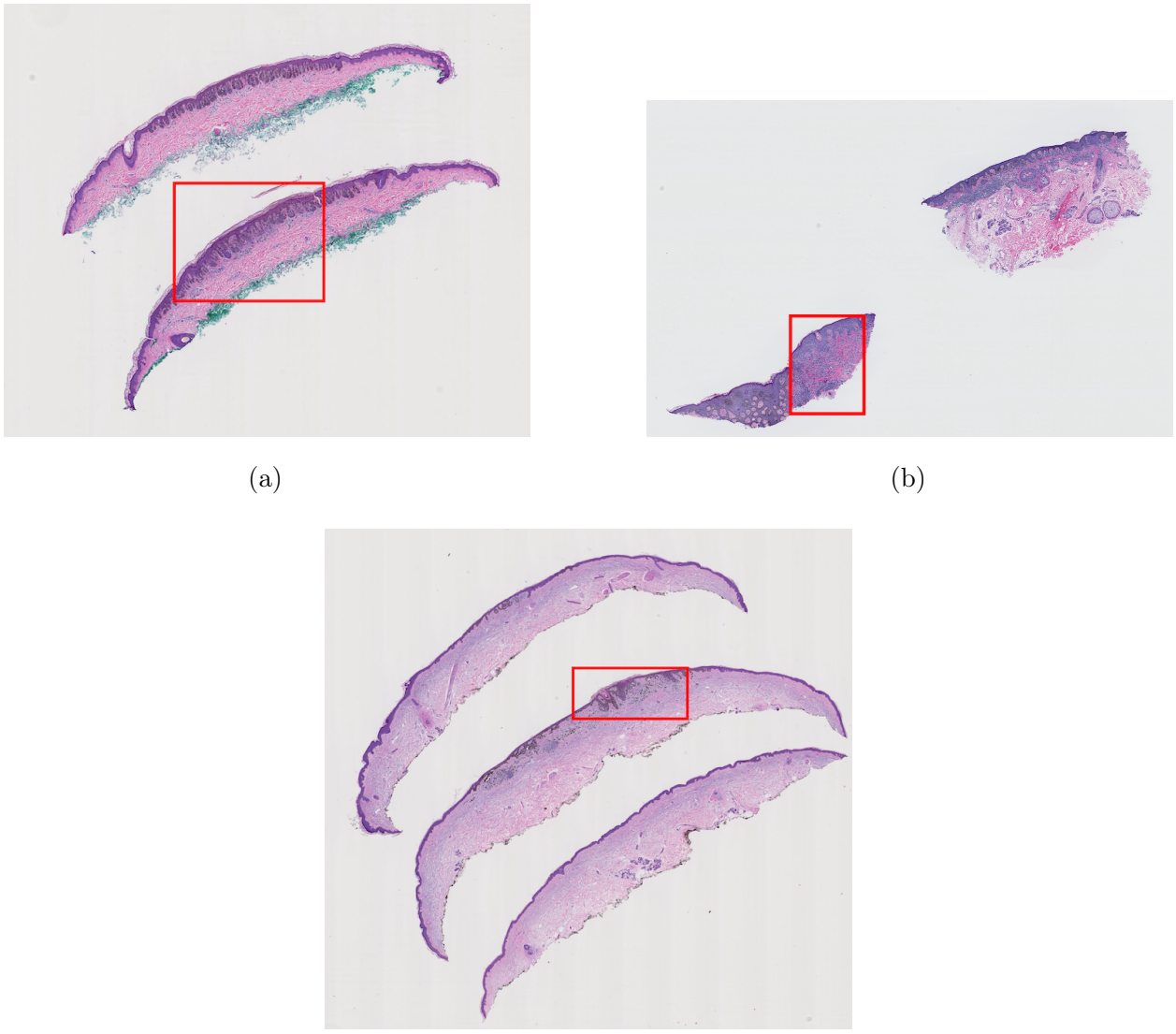
Figure 2.1: Three examples of WSIs in the M-Path dataset

A consensus panel of three dermatopathologists with internationally recognized expertise met over several days to reach consensus diagnoses for all cases, using the aforementioned

Table 2.1: Distribution of diagnostic categories in M-Path data

Diagnostic Category	**#Cases**
Class I (e.g. Mildly Dysplastic Nevi)	25
Class II (e.g. Moderately Dysplastic Nevi)	36
Class III (e.g. Melanoma in Situ)	60
Class IV (e.g. Invasive Melanoma Stage T1a)	72
Class V (e.g. Invasive Melanoma Stage T1b) $\geq$ T1b	47
Total	240

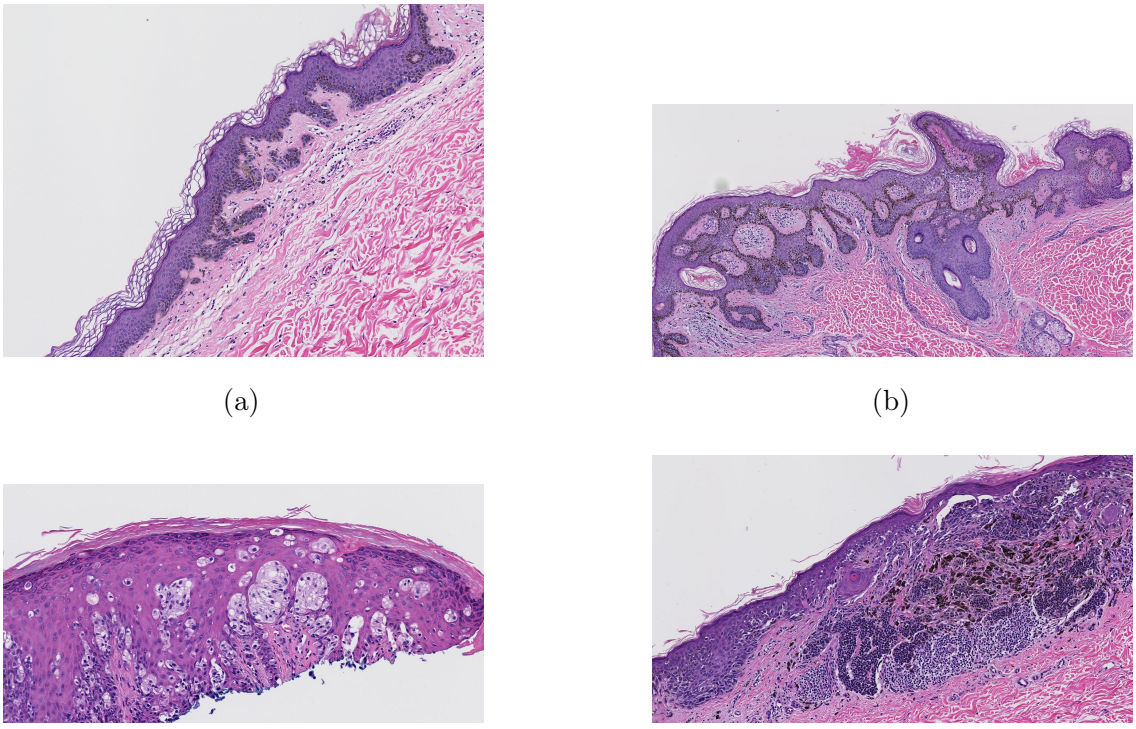
MPATH-Dx classification tool [5] explained in . Following these consensus meetings, the consensus panel members, as well as an additional dermatopathologist on the M-Path research team (S. Knezevich), utilized digitized images of all cases to identify one rectangular area as a Region of Interest (ROI) per case. These regions represent an important area of the WSI for the diagnosis. The size of these ROIs is not fixed and varies from one case to another (Figure 2.2). We can extract the ROIs using their coordinates and perform various analyses on them to improve the overall diagnosis (Figure 2.3).



(a) (b)

(c)

Figure 2.2: Examples of variable-sized whole slide images are shown. Regions of interest (ROIs) that helped pathologists in diagnosis are shown in red boxes.



(a) (b)

(c) (d)

Figure 2.3: Examples of four variable-sized ROIs with different diagnoses in the M-PATH dataset. **(a)** MPATH-Dx Class I: Mildly Nevus. **(b)** MPATH-Dx Class II: Moderately Nevus. **(c)** MPATH-Dx Class III: Melanoma in-Situ. **(d)** MPATH-Dx Class IV: Invasive Melanoma Stage T1a.

Chapter 3

MITOSIS CLASSIFICATION

3.1 Introduction

Melanoma diagnosis involves histological analysis of various cellular and architectural features. Melanocytic lesions range across a broad spectrum of categories: I) Mildly Nevus, II) Moderately Nevus, III) Melanoma in Situ, IV) Invasive Melanoma Stage T1a and V) Invasive Melanoma Stage $\geq$ T1b. [65]. A mitosis (or mitotic figure) remains an important entity in the review of skin biopsy cases as its presence may aid in the diagnosis of a melanoma in addition to being associated with poorer prognosis. A high mitotic rate in a primary invasive melanoma is associated with a lower survival probability. Among the independent predictors of melanoma-specific survival, mitotic rate is the strongest prognostic factor after tumor thickness [81]. Thus, the accurate detection of mitotic activity is an important role for the pathologist in making cancer diagnoses, and because mitoses are small objects with various shapes that can resemble normal nuclei, mitosis detection remains a challenging task for humans. Because of its clinical importance, the development of automated mitosis detection has become an active area of research with the goal of developing decision support systems to assist pathologists [40].

Various approaches have been applied to detect mitotic figures. [76] computed the probability map based on the likelihood functions and then used a component-wise two-step thresholding to find mitoses in neuroblastoma. A graph-based multi-resolution approach with color and texture features was used by [72] for mitosis extraction in breast biopsy images. Irshad et al. used morphological features to identify cellular entities in a breast biopsy dataset [32].

In recent years, with the development of fast and accessible Graphics Processing Units

(GPUs), Convolutional Neural Networks (CNNs) have gained attention for medical image analysis, primarily because of their capability to learn strong structural representations about objects of interest (e.g. cellular entities [11] or tissues [49, 70]). For example, [11] used a CNN-based method for mitosis detection and won the International Conference on Pattern Recognition 2012 (ICPR 2012) mitosis detection challenge by a significant margin. Since then, much of the research on mitosis detection in breast cancer biopsy images has used CNNs. [77, 32] and [88] developed different methods that merge CNN image descriptors and handcrafted features to improve the detection. [7] proposed a two-stage mitosis detection pipeline, with a coarse retrieval model, followed by a fine discrimination model. In recent work, [40] used a deep detection network using residual connection when only the weak label is given. [42] introduced a pyramidal model to detect mitoses. On each pyramid level, a Bayesian convolutional neural network is trained to compute class prediction and uncertainty on each pixel.

Several CNN-based methods have been proposed for mitosis detection in different tissues, including breast [11, 32, 7], stem cells [100], and skin [42]. Unlike natural image datasets (e.g. the ImageNet [17]), the number of training samples are limited in medical image datasets usually by an order of a few hundred ([71, 84, 83]. To achieve strong performance on these datasets, CNNs have been complemented with several methods, including handcrafted features ([73, 32, 18]) and better augmentation strategies [70]. U-Net [70] introduced an encoder-decoder architecture with skip-connections for segmenting different biological structures in images and demonstrated good performance across several datasets.

Most research in mitosis detection has been conducted on biopsy images other than the skin ([73, 53, 49, 7]). However, skin biopsy images are different from these biopsy images in terms of texture, color, and mitosis shape, as shown in Figure 3.1. As a result, existing CNN-based classifiers trained on these biopsy images may have poor performance on skin biopsies. Moreover, to the best of our knowledge, there are no publicly available skin biopsy datasets with mitosis annotations. Given the importance of mitosis detection in skin cancer diagnosis, we created a new dataset with mitosis-level markings from an expert pathologist.

We studied and compared the performance of two different state-of-the-art CNNs, one that is lightweight in terms of parameters and execution time and one that is much bigger, in terms of accuracy, sensitivity, specificity, precision, recall, and F-score. We then compared the performance of these two CNNs with two additional state-of-the-art architectures on a public breast cancer data set in terms of precision, recall, F-score, architecture, training time, and inference time. Our work has several contributions: 1) This is the first work to experiment with finding mitotic figures in whole slide melanoma biopsies. 2) After determining the best possible performances on the melanoma biopsy slide images, we showed that this pipeline could be applied to a well-known breast cancer data set (MITOS) and compared the results from our two models (ESPNet, which was chosen for lightweight network and speed, and DenseNet, which was an example of a state-of-the-art network) with the results from several published papers, showing that DenseNet could beat all of them and ESPNet came close (Table 3.4). 3) We ran two more models, ResNet and ShuffleNet, on the MITOS dataset for further comparison and found that DenseNet is still the best performer in terms of F-1 score (DenseNet 0.927, ESPNet 0.890, ResNet 0.865 and ShuffleNet 0.847) and, particularly, in terms of Recall (DenseNet 0.916, ESPNet 0.866, ResNet 0.870 and ShuffleNet 0.753), which is very important for cancer grading. 4) Our research, in general, gives a methodology and architecture for mitosis finding in both melanoma and breast cancer whole slide images, and that is likely to be useful for finding mitoses in any whole slide biopsy images.

3.2 Dataset

An experienced pathologist (S. Knezevich) chose six skin biopsy cases of $\geq$ T1b invasive melanoma, a diagnostic category known to be associated with high mitotic activity, from our dataset and cropped 34 areas in the whole slide images (WSIs) of these cases. The size of the areas and the number of areas per each case were not fixed but were based on the pathologist's judgment with the aim of marking as many mitoses as possible. A total of 628 mitoses in the cropped image areas were marked by the same pathologist with a green dot on each mitosis, using the Sedeen Viewer. These marked mitoses provide "class mitosis" samples

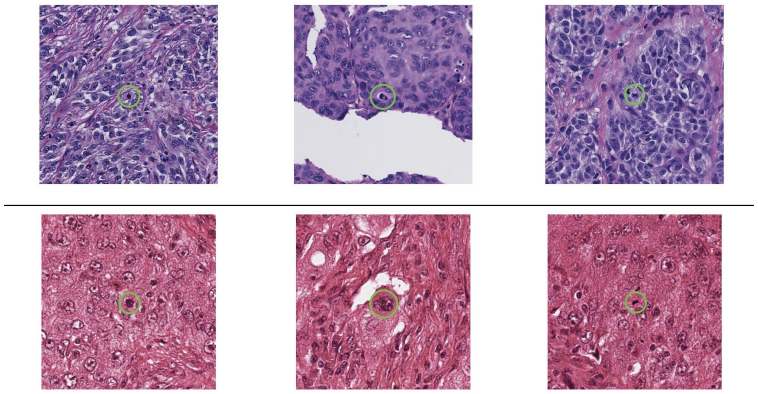


Figure 3.1: Example crops of biopsy images with mitoses in them; (top) skin; (bottom) breast. These biopsies are different in terms of color, texture, and mitosis phase and shape. *A mitosis in each image is present near the center and is marked with a green circle for visualization.*

for training and validation of our binary classifiers. The details about our skin biopsy dataset are summarized in Table 3.1.

Distinguishing mitoses from normal nuclei is a challenge for automated mitosis classifiers. Mitoses and nuclei can appear very similar in color and shape; thus, the classifiers require a large number of nuclei samples to differentiate between these cellular entities. If the whole non-mitosis regions of the image were to be sampled uniformly, many of the non-interesting instances such as background would be in the class "non-mitosis" and training a strong classifier would be inefficient. To avoid this, we used a standard watershed-based nuclei segmentation method [13] to find nuclei in the images and used them as examples for the class non-mitosis. Figure 3.2 shows the output of this nuclei detector on a cropped portion of a skin biopsy.

Figure 3.3 shows some examples of mitoses and normal nuclei, which we note are very similar in terms of texture, color, and shape. In the process of sampling mitoses and nuclei, based on our experiments, we used a 101×101 patch approximately centered on the target entity's center. If a part of this window lies outside of the image borders, the image is padded using mirroring of the border pixels. To help our classifier learn rotation, scale,

Table 3.1: Mitosis dataset summary – **Melanoma**.

Case ID	**#slices**	**#cells in WSI**	**#areas**	**#mitosis**
Case #1	5	$\sim$ 250k	14	197
Case #2	3	$\sim$ 237k	6	32
Case #3	6	$\sim$ 320k	7	232
Case #4	1	$\sim$ 115k	5	156
Case #5	3	$\sim$ 49k	1	6
Case #1	4	$\sim$ 39k	1	5
Total	-	-	34	628

and translation-invariant representations, we augmented our training set with standard augmentation methods such as rotation (45, 90, 135 or 225 degrees) and mirroring (horizontal and vertical).

The number of mitoses per slide is an order of magnitude fewer than other entities, such as nuclei and melanocytes present in the slide. In other words, the dataset is imbalanced. If we train a classifier with such an imbalanced dataset, then the classifier will be biased towards the entities with more samples. To address this imbalance, a standard approach [66, 67] is to maintain a good ratio between positive samples (patches that contain mitoses) and negative samples (patches that do not contain mitoses). For our dataset, we empirically found that this ratio is 1:3 i.e. the number of negative samples available for training is approximately 3 times the number of positive samples; resulting in 4364 mitoses and 12,640 non-mitosis samples after data augmentation. Since we used a watershed-based nuclei segmentation [13] as a pre-processing method, non-mitosis samples mostly contain nuclei.

We split our dataset randomly into training (80 %) and validation (20 %) sets, respectively. The validation set was withheld during the training phase. After the training was complete,

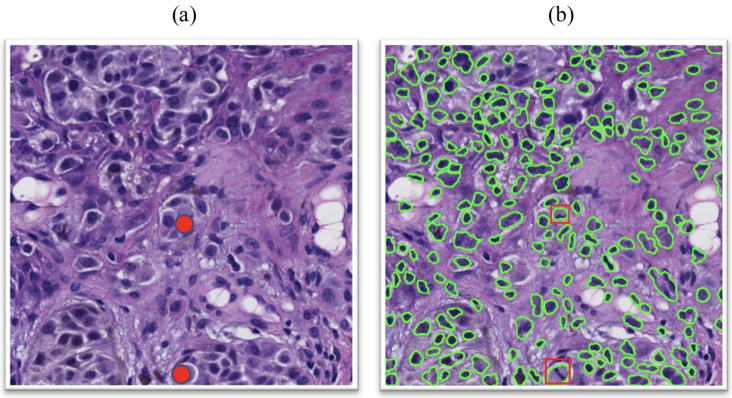


Figure 3.2: Examples of applying the nuclei segmentation method [13] on a crop of skin biopsy image (a) original crop (b) nuclei segmentation result. Two mitoses that are present in the original crop are marked with red dots for visualization. *Segmentation method was able to find the mitoses. We marked them here with red boxes for visualization.*

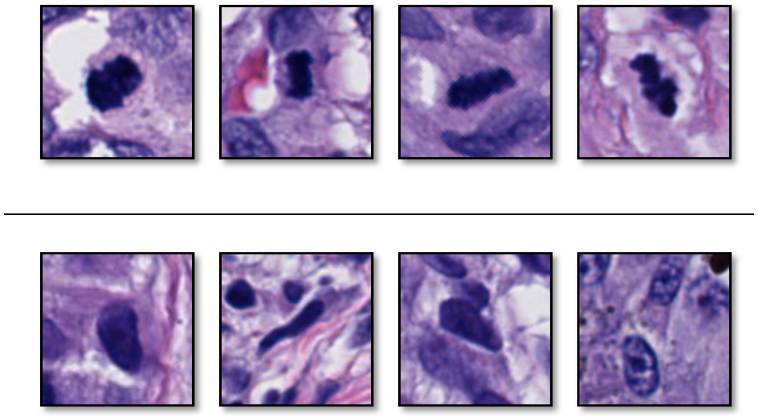


Figure 3.3: Examples of (**top**) sampled mitoses, and (**bottom**) sampled nuclei that are not mitoses. These two entities have similarity in color, surrounding and texture.

the validation set was used to evaluate the trained model performance.

3.3 Method and Model

3.3.1 Training and Inference

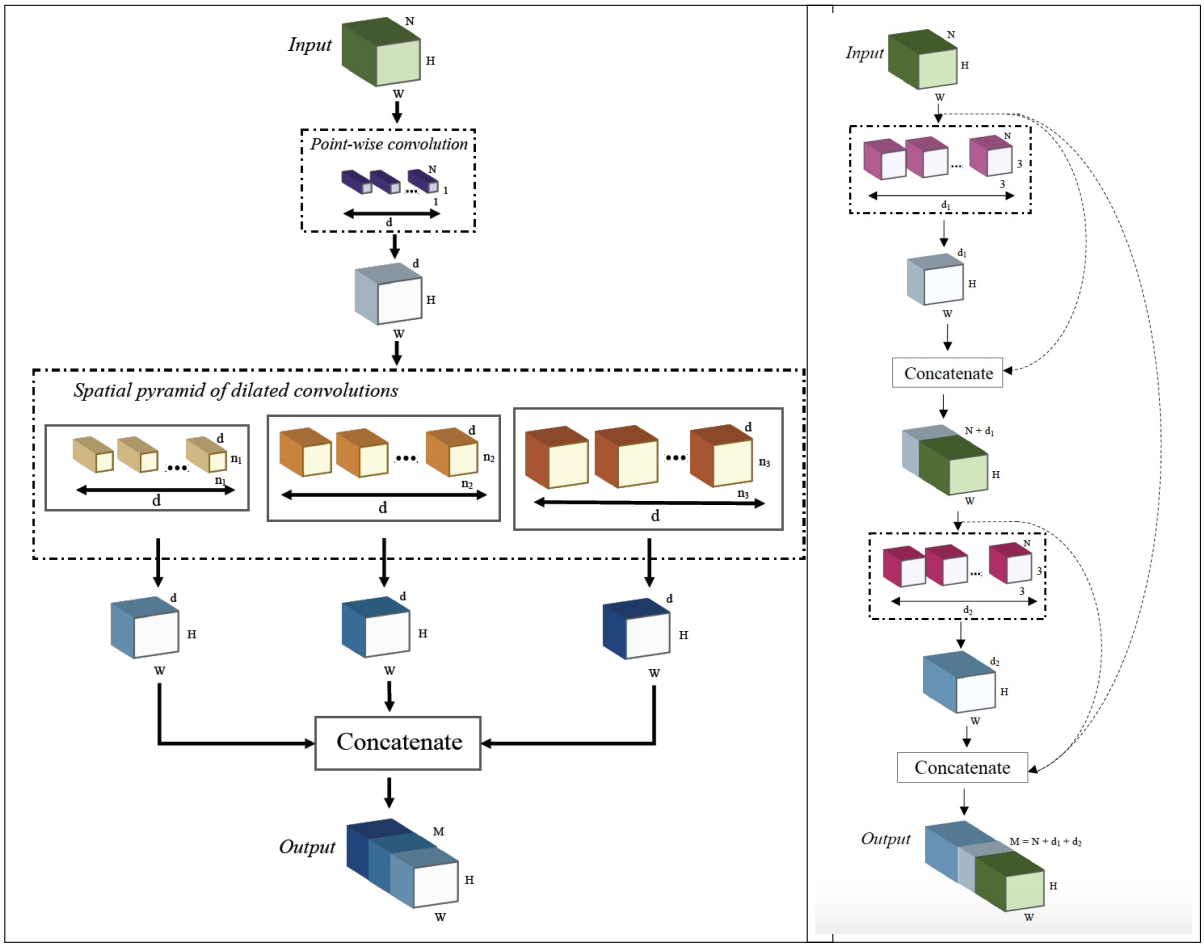
Our classification network uses a standard pipeline [37, 28] that stacks encoding and down-sampling units to learn latent representations. In our experiments, we used two state-of-the-art encoding units: 1) Efficient Spatial Pyramid of Dilated Convolutions (ESPNet) [50] and 2) Densely Connected Convolutional Networks (DenseNet) [31]. The same dataset split was used for both ESPNet and DenseNet training and validation.

Efficient spatial pyramid of dilated convolutions (ESPNet): ESPNet [50] is a fast and efficient CNN that was designed for semantic segmentation on mobile devices. The core building block of the ESPNet architecture is the ESP unit that decomposes a standard convolution into a point-wise convolution and a spatial pyramid of dilated convolution. This factorization reduces the computational complexity of the ESP unit in comparison to the standard convolution. Figure 3.4a visualizes the ESP unit. We chose this unit in our study because of its good performance in segmenting breast biopsy whole slide images[49].

Densely Connected Convolutional Networks (DenseNet): DenseNet, densely connected convolutional neural network [31], introduces a novel connectivity mechanism to improve the flow of information between different stacked convolutional layers. As shown in Figure 3.4b, this unit establishes a direct link between different convolutional layers. This connectivity pattern provides multiple paths for gradients to flow back to the input and thus, helps in learning better representations.

3.3.2 Hyperparameters and Loss Function

We trained our classifiers using the ADAM optimizer [34] for a total of 20 epochs with an initial learning rate of 0.001. We decayed the learning rate by 0.1 after every 5 epochs. During training, we minimized the cross-entropy loss [16].



(a) ESPNet unit (b) DenseNet unit

Figure 3.4: Two convolutional units, ESPNet (a) and DenseNet (b), that are used in our experiment. Each of these units receives a 3D tensor with width W, height H, and depth N as an input and produces a 3D tensor with width W, height H, and depth M as an output. The projection channel dimension in ESPNet unit is represented by d while in DenseNet unit, it is represented by d_i. For ESPNet, output tensor depth is M = k × d, where k is the number of parallel branches in the ESPNet unit (k = 3 in (a)), the size of the point-wise convolution is 1 × 1, and n_i is the size of the dilated convolutional layers. For more information, see [50]. For the DenseNet unit, output tensor depth is M = $\sum d_i$, i = 1,. . . , L, where L represents the number of stacked layers (L = 3 in (b)). It is common to use 3 × 3 standard convolutional layers in DenseNets. For more information, see [31].

3.4 Results

3.4.1 Evaluation Metrics

We evaluated the performance of our classifier on the melanoma dataset using six metrics: four standard metrics (precision, recall, F-score, and accuracy) and two widely used metrics in clinical care (sensitivity and specificity):

- $Accuracy = (TP + TN)/(TP + FP + FN + TN)$
- $Precision = TP/(TP + FP)$
- $Recall = TP/(TP + FN)$
- $F1\ score = 2 \times \frac{Precision.Recall}{Precision+Recall}$
- $Sensitivity = TP/(TP + FN)$
- $Specificity = TN/(TN + FP)$

where TP is the number of true positives and FP the number of false positives; TN is the number of true negatives and FN the number of false negatives; *F1 score* is the harmonic mean of precision and recall.

3.4.2 Mitosis detection results on Melanoma dataset

Table 3.2 summarizes the results of our classifiers using two different encoding units: 1) ESPNet and 2) DenseNet. Both networks achieved high accuracy on classifying mitoses with a sensitivity of 0.976 and 0.968, and specificity of 0.987 and 0.995, respectively. Though DenseNet outperformed ESPNet, this outperformance was not statistically significant (p-value is 0.5716), and the training time of ESPNet is about a third that of DenseNet (see Table 3.2) (Table 3.3).

Table 3.2: Quantitative results of ESPNet and DenseNet on validation set **Melanoma**

Metrics	**ESPNet**	**DenseNet**
Accuracy	0.984	0.988
Precision	0.961	0.984
Recall	0.976	0.968
F1 Score	0.968	0.976
Sensitivity	0.976	0.968
Specificity	0.987	0.995
FP, FN	5, 3	2, 4
TP, TN	122, 370	121 , 373
Training time	35m & 6s	106m & 32s

3.4.3 Generalizability to the MITOS dataset

To study the generalization ability of our classifiers on other datasets, we evaluated the performance on a publicly available mitosis dataset for breast biopsies: MITOS [71]. The dataset consists of 50 images corresponding to 50 high-power fields in 5 different breast cancer slides stained with hematoxylin and eosin. We first compared our two classifiers (ESPNet and DenseNet) to the results reported in several papers in the recent literature [73, 11, 40, 42, 18]. The architectures of these classifiers can be summarized as follows:

- **Saha, et al.** The deep learning consists of two parts: (1) a convolutional neural network and (2) a handcrafted feature extractor. The deep architecture contains five convolution layers, four max-pooling layers, four ReLUs, and two fully connected layers.

- **Dodballapur et al.** In this work, handcrafted features extracted from the masks generated from the Mask R-CNN network are combined with deep features to classify

the candidate cells. To extract an image-level representation, the Xception network pre-trained on ImageNet without the last two fully connected layers was used.

- **Li, et al.** Their pipeline consists of three components: (1) a deep detection model (DeepDet) that produces primary detection results, (2) a deep verification model (DeepVer) that verifies these detections and eliminates false positives, and (3) a deep segmentation model (DeepSeg) that segment the images and generates bounding box annotations around segmented regions to provide weak box-level annotations. The DeepDet model consists of an RPN (Region Proposal Network) and a region-based classifier. The DeepVer model is based on the ResNet.

- **L´opez-Tapia, et al.** Their pipeline consists of two components: first, a coarse-to-fine cascade of CNN Bayesian models for mitosis detection; then, to make the model resistant to local and shape deformations, a Spatial Transforming Layer is applied before the 4th and 7th residual blocks in scale x40.

- **Cires¸an, et al.** They trained two DNNs and ensembled the performance evaluation results: DNN1 contains five convolutional layers, five max-pooling layers, and two fully connected layers. DNN2 contains four convolutional layers, four max-pooling layers, and two fully connected layers.

For comparison, the architectures of ESPNet and DenseNet are as follows:

- **ESPNet:** Our classification network uses a standard pipeline that stacks encoding and down-sampling units to learn latent representations. The model contains one conventional 2D convolution layer, five ESP blocks, four down-sampling layers, one average-pooling, and two fully connected layers.

- **DenseNet:** We used the DenseNet161 architecture which contains one conventional 2D convolution layer, four Dense block, three Transition layers, one max-pooling, and two fully connected layers.

Table 3.3: Quantitative results of ESPNet and DenseNet on **MITOS** [71]

Metrics	**ESPNet**	**DenseNet**
Accuracy	0.946	0.964
Precision	0.916	0.939
Recall	0.866	0.916
F1 Score	0.891	0.927
Sensitivity	0.866	0.916
Specificity	0.973	0.980
FP, FN	16, 27	12, 17
TP, TN	175, 586	185 , 586

In comparison to existing state-of-the-art methods (see Table 3.4), our classifiers achieve a competitive performance. In particular, our DenseNet-based classifier is 2% more accurate than [73].

In order to compare more thoroughly, we added two more state-ofthe-art CNNs, ResNet [28] and ShuffleNet [98] to the original two (ESPNet and DenseNet). We compared all four classifiers on precision, recall, and F-score (as is standard for MITOS) and measures of architecture and speed.

Results with precision, recall and F-score are summarized in Table 3.5. DenseNet is the clear winner in this contest with F-score of 0.927 compared to 0.890 for ESPNet, 0.865 for ResNet and 0.847 for ShuffleNet. Furthermore, results with respect to architecture and speed are summarized in Table 3.6. Here ResNet is the most efficient with ESPNet a close second.

Table 3.4: Performance comparison of ESPNet and DenseNet with other approaches on MITOS [71] reported in the literature.

Method	**ESPNet** (Ours)	**DenseNet** (Ours)	[73]	[18]	[40]	[42]	[11]
Precision	0.916	**0.939**	0.92	0.93	0.854	N/A	0.866
Recall	0.866	**0.916**	0.88	0.80	0.812	N/A	0.70
F1 Score	0.890	**0.927**	0.90	0.87	0.832	0.826	0.782

Table 3.5: Quantitative results of ESPNet, DenseNet, ResNet, and ShuffleNet on **MITOS** [71]

Metrics	**ESPNet**	**DenseNet**	**ResNet**	**ShuffleNet**
Precision	0.916	0.939	0.931	0.968
Recall	0.866	0.916	0.807	0.753
F1 Score	0.891	0.927	0.865	0.847

3.5 Summary

One microscopic parameter that is both helpful to the pathologist in establishing a cancer diagnosis and in assessing prognosis, is the presence or absence of mitotic figures; a microscopically visible nuclear feature closely tied to cellular proliferation. In mitosis a cell divides to form two new cells. Cancer tissue generally has more mitotic activity than normal tissues, and this is assessed by calculation of the mitotic index – the number of cells in mitosis divided by the total number of cells However, measurement of the mitotic index depends on the subjective visual analysis by pathologists who have a hard time both in identifying and also counting mitotic figures and total cell counts [35]. Thus, development of supporting tools

Table 3.6: Architecture, training and inference time comparison of ESPNet, DenseNet, ResNet, and ShuffleNet on **MITOS** [71]

Network	**#parameters** (in million)	**#block** (depth)	**# channels** (width)	**Training time**	**Inference time**
ESPNet	0.078	16	16 to 64	6 min	8 sec
DenseNet	28.68	161	48 to 2024	19 min	31 sec
ResNet	11.69	12	64 to 512	4 min	6 sec
ShuffleNet	2.28	56	24 to 1024	6 min	11 sec

that can be more accurate and reproducible would greatly aid clinical care. Machine learning techniques, including CNNs, have shown incredible performance in visual recognition tasks, and thus have the potential to improve histologic diagnostics, both as aids for pathologists to improve the quality and reproducibility of their diagnoses and in the medical research domain [49, 68, 33].

In this work, we trained two CNN methods, ESPNet and DenseNet, as two separate classifiers; both CNNs had high accuracy on our dataset of skin biopsies of invasive melanoma. We further generalized our classifiers to the MITOS breast biopsy dataset and compared our results with the existing state-of-the-art on the MITOS dataset with high accuracy in classifying mitoses [73, 11, 7, 40, 42, 18] and ran experiments with two more state-of-the-art CNNs to make more thorough comparisons. We achieved competitive accuracy on the MITOS dataset compared to the existing state-of-the-art methods.

No study is without limitations, and our research is not an exception. First, both the melanoma dataset and the MITOS dataset (as well as other public digital datasets) make use of less information than a microscopic examination, in which a typical tissue section is 5 μm and on which the pathologist can focus through an infinite number of planes, ensuring all cells of interest are in optimal focus. Secondly, for the public datasets, the use of only

two-dimensional images with no recourse to looking at three-dimensional tissue sections makes it difficult to confirm the given diagnoses.

Marking biopsy images is an onerous task and obtaining samples with variation in the dataset is a challenge. To expand our dataset, we generated new samples out of our existing samples with horizontal and vertical mirroring and with rotations of 45, 90, 135 or 225 degrees. However, having samples from more patients would be beneficial for training a precise classifier for mitosis detection.

Given the complex and dense nature of working with biopsy tissue datasets, a significant challenge is posed in developing training sets that reflect the full spectrum of cases seen in clinical practice and also that accurately identify the cellular entity of interest. In our skin cancer work, the cases were carefully selected to represent the full spectrum of skin biopsies obtained in clinical practice and a three-person expert defined consensus diagnosis was used [19]. In addition, each case was carefully reviewed by an expert dermatopathologist to identify and mark the individual mitotic figures.

Mitotic activity is an important biomarker that can assist in the diagnosis and may provide prognostic information. However, each biopsy specimen may contain hundreds of thousands of cells, making their identification a significant challenge. We have shown that mitoses can be identified using our machine learning method with high accuracy; thus, this method has the potential of being a powerful diagnostic and prognostic aid to practicing pathologists.

Chapter 4

SEGMENTATION USING IMPERFECT ANNOTATIONS

4.1 Introduction

The histologic evaluation of melanocytic lesions, including melanoma and its precursors, involves determining whether the melanocytic population involves the epidermis, dermis, or both. For example, the atypical melanocytes in melanoma in situ are contained within the epidermis, whereas an invasive melanoma shows atypical melanocytes which in the the dermis. Semantic segmentation of various structures in skin biopsy images, including accurately distinguishing between the epidermis/dermis and identifying epidermal/dermal melanocytes, has the potential to improve the automated diagnosis systems or serve as a diagnostic aid in the decision-making process. The goal of semantic segmentation is to label each pixel of an image with the corresponding class of the objects being represented. Hence, semantic segmentation of clinically-relevant structures in skin biopsy images can play a key role in an automated diagnosis system.

One key challenge in training a segmentation model is that it requires large-scale and fine annotations. However, collecting fine tissue-level annotations for biopsy images is an onerous, exhaustive, and expensive task because of the sheer size of biopsy images and the fact that domain experts are required for annotations. As a result, full annotation of the whole slide image (WSI) for large datasets is the leading limitation of medical imaging research. This work introduces a simple two-step approach for learning representations with coarse and sparse labels. An overview of our approach is shown in Figure 4.1. The core principle is to segment larger and smaller entities separately, allowing us to segment images with good accuracy.

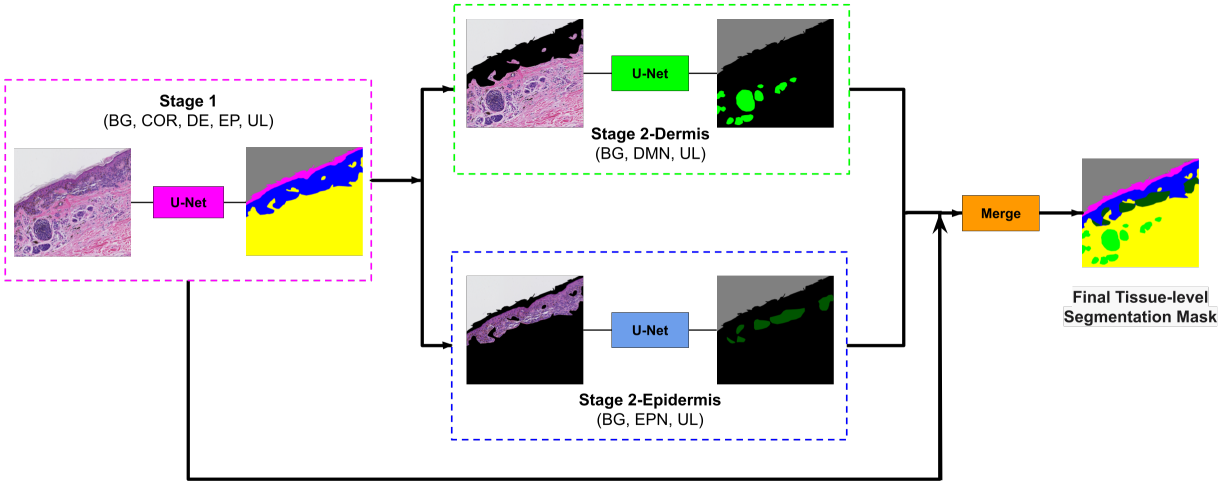


Figure 4.1: **Overview of our approach**. The image first goes to stage 1 and the segmentation mask of entities (COR: Stratum Corneum, EP: Epidermis, DE: Dermis, BG: Background, and UL: Unlabeled) in stage 1 is generated. Then this mask is used to remove the epidermis from stage 2-Dermis input and remove the dermis from stage 2-Epidermis input. The modified images are fed to their corresponding trained model. Stage 2-Dermis generates the segmentation masks of entities present in the dermis (DMN: Dermal nests), and stage 2-Epidermis generates the entities in the epidermis (EPN: Epidermal nests). In the end, stage 2-Dermis and stage 2-Epidermis segmentation masks are overlaid on the stage 1 mask and the final tissue-level segmentation mask is generated.

4.1.1 Related Work

Various approaches have been developed to overcome imperfect and limited data annotation and vary with the specific challenges posed by the specific dataset on which they were developed. When a small portion of an image is fully annotated, different methods of augmentation have proven to be helpful. [96] showed that data augmentation by adjusting image quality produces performance gain in magnetic resonance imaging (MRI), especially image sharpening through the application of unsharp masking, which has the largest improvement. In another study, [41] proposed asymmetric mixup that turns soft labels generated by mixup into hard labels, which improves the segmentation of brain tumors according to their experiments.

Active learning is another popular method in the case of limited annotation. [79] proposed a probabilistic active learning pipeline where the probability of an unlabeled sample that is queried in the next round of annotation is estimated based on its Fisher information. [46] used a Bayesian neural network for active learning: using a combined metric based on noise in the data and uncertainty over their Convolutional Neural Network (CNN) parameters, they selected the most informative samples. [99] proposed a one-shot active learning method, which eliminates the need for iterative sample selection and annotation. However, active learning generally requires a base segmentation model with careful annotation; hence, a dataset with only coarse annotations may not benefit from active learning, unless a pre-trained model from a similar domain is available [80].

In some studies, modification of a loss function solved the sparse annotation challenge to some extent. [3] used class-balancing methods to improve the segmentation performance given sparse annotations without trying to fill in the missing mask pixels. In this proposed method, only the labeled pixels contribute to a weighted segmentation loss. The dataset used in this work contains some densely annotated WSIs and some sparsely annotated WSIs. However, segmenting whole slides images using coarse and sparse annotations is challenging and remains understudied in the literature.

Utilizing domain adaptation and leveraging external data have generated promising

segmentation results. However, to the best of the authors' knowledge, no carefully-labeled public dataset is available on skin biopsy images, and datasets from other domains have significantly different morphological features compared to those of skin biopsy images.

There are limited studies on skin biopsy image segmentation. [94] presented a robust technique for epidermis segmentation in whole slide skin histopathological images, using thresholding and shape analysis. [62] produced a model for segmenting psoriasis-affected human skin biopsy images into the dermis, epidermis, and non-tissue regions. [2] developed a fully automated technique for lymph node segmentation that is robust to stains such as H&E, MART-1, S-100, KI-67. However, semantic segmentation of clinically important structures in skin biopsy images is one of the most understudied areas in the literature. This is especially true for the datasets with imperfect and limited ground-truth annotations.

In this chapter, we describe a carefully designed segmentation pipeline that can train a CNN on images with coarse and sparse annotation to accurately segment clinically important tissue structures in WSIs. This approach can be extremely helpful for medical image researchers because both data and annotations are expensive to acquire.

4.2 Dataset

4.2.1 Coarse and Sparse Segmentation Annotations

To train a segmentation model, labels of different tissues as the ground-truth are required. However, since the annotation task is a very labor-intensive task, we obtained coarse and sparse annotations only on the ROI images by an expert pathologist (M. Mokhtari). Not only are the annotations not on the full WSI (Figure 4.2a), but they are also sparse within the annotated ROI c. Moreover, the annotations are coarse, i.e., they are not pixel-level accurate, as shown in Figure 4.2c.

For pixel-level annotations, Sedeen[1], a pathology image viewer, was used. Various structures were labeled with different names and corresponding colors as follows: Epidermis

[1] `https://pathcore.com/sedeen`

(EP) in blue, Dermis (DE) in yellow, Stratum Corneum (COR) in pink, Epidermal Nests (EPN - corresponding to epidermal melanocytic nests) in dark green, Dermal Nests (DMN - corresponding to dermal melanocytic nests) in light green. Using the threshold-based segmentation method of [61], the background (BG) was detected and added to the labels in gray. We followed existing segmentation dataset annotation protocols (e.g., Cityscapes) and marked the pixels that do not correspond to any informative entity as unlabeled (UL) in black.

4.3 Method and Model

Medical imaging literature has witnessed great progress in the design and performance of deep convolutional models for medical image segmentation [80]. Thus, we utilized a CNN for our task of semantic segmentation. Since the labeling was done on ROIs, we started the process of training and evaluation on ROIs. As the preprocessing step, cropping, resizing, and augmentation was performed on these images.

4.3.1 Cropping and Resizing

Since the ROIs sizes vary from ∼(480 x 360) to ∼(23500 x 22400), we chose the smallest size of (480 x 360) as the model input. However, resizing the biggest crop of (23500 x 22400) to such a small size (480 x 360) can significantly impact the information that can be acquired from such images. Instead, we follow a standard approach wherein bigger ROIs are divided into patches and then patch-level segmentation masks are generated and combined to produce a ROI-level segmentation mask [49, 70]. In particular, for bigger ROIs, we extract patches of size 1440×1080 and then resize them to 480×360 before feeding them to the model. The segmentation output is then upsampled using nearest-neighbor interpolation to produce the segmentation mask that is of the same size as the patch before resizing.

4.3.2 Augmentation

We used various augmentation techniques such as horizontal flipping, affine transformation, perspective transformation, brightness/contrast/color manipulations, image blurring, sharpening, Gaussian noise, and random cropping to improve the robustness of our model. We used the fast augmentation library for these augmentation techniques [4].

4.3.3 Data Split

For a fair evaluation of the model, we divided the ROI dataset into two subsets, the training and testing sets, with a ratio of 80/20, respectively. The testing set was kept unseen from the model until the last step of the final evaluation. The training set is further split into train and validation sets, with a ratio of 80/20. We use the validation set for monitoring the training process and model selection , and the testing set for the evaluation.

4.3.4 U-Net

The U-Net architecture of [70] is a well-known segmentation network and has shown good performance across different biomedical segmentation applications; such as MRI images [10], COVID-19 [75], skin lesion images [55], and lung, heart, and clavicl X-Ray images [22].

We extended the U-Net encoder-decoder model for segmenting skin biopsy images. We used the implementation of U-Net by [95] in our work. We used ResNet-34 [27] pre-trained on the ImageNet dataset [17] as the encoder and a standard U-Net decoder [95] [2].

4.3.5 Two-Stage Pipeline

Skin biopsy images have entities of variable size. Entities like the dermis and epidermis are large and easy to segment [94], while entities like dermal and epidermal nests are small and more difficult to segment. This issue becomes especially more troublesome when the smaller entities have sparser labeling compared to the larger entities. Hence, if the segmentation

[2]We did not use Attention in the U-Net decoder block

model is trained in a single-stage with all the labels at once, the model will perform better on larger entities and not as well on the smaller ones.

To overcome this problem, we developed a two-stage segmentation pipeline: First, a segmentation U-Net model is trained with labels of large entities in the histopathology image (Background, Stratum Corneum, Epidermis, Dermis). Then, in the second stage, there are two sub-stages: 1) Stage 2-Dermis is trained on the dermis portion of the images and uses the ground truth for the smaller entities that are present in Dermis (i.e., DMN). 2) Stage 2-Epidermis is trained on the epidermis portion of the images and uses the ground truth for the smaller entities that are present in Epidermis (i.e., EPN). [3] Figure 4.3 shows an example of one ROI and its corresponding mask, which is modified for different stages of our proposed pipeline.

As previously mentioned, the whole segmentation pipeline is trained in two stages: the first stage for big entities, such as dermis and epidermis, and the second stage for smaller entities within the dermis and epidermis, such as dermal nests and epidermal nests. All the training stages used the Stochastic Gradient Decent (SGD) optimizer with a momentum of 0.9, and a learning rate of 0.0001. The stage 1 encoder is trained for 1000 epochs. For the second stage weight initialization, U-Net in both the Dermis and Epidermis branches was initialized with the stage-1 model and fine-tuned for 100 epochs. This helps the model to converge faster. All the experiments were performed on an Intel(R) Xeon(R) Silver 4110 CPU 2.10GHz with NVIDIA GeForce GTX 1080 GPU.

4.3.6 Evaluation Metrics

To evaluate our models, we used mean Intersection over Union (IoU). IoU is a number from 0 to 1 that specifies the amount of overlap with the ground-truth (Equation 4.1). An IoU of 0 means that there is no overlap between the prediction and ground-truth and an IoU of 1 means the prediction and ground-truth completely overlap. Thus, a higher value of IoU

[3] While there are other small structures, such as hair follicles and blood vessels present in skin biopsy images, they are not clinically important for the diagnosis, so we do not try to segment them in this work.

means better performance.

$$IoU = \frac{TP}{TP + FN + FP} \tag{4.1}$$

For the final evaluation, we calculated another metric, Dice Coefficient which is 2 * the area of overlap divided by the total number of pixels in both images (Equation 4.2).

$$Dice = \frac{2 * TP}{(TP + FP) + (TP + FN)} \tag{4.2}$$

where True Positive (TP) is the number of pixels that are correctly predicted as nest, True Negative (TN) is the number of pixels that are correctly predicted as not-nest, False Negative (FN) is the number of pixels that are incorrectly predicted as not-nest, and False Positive (FP) is the number of pixels that are incorrectly predicted as nest.

Acquiring pathologists' annotations was a challenge. While we did not have full annotations for the whole dataset, we acquired fine-grained nest annotations on ROIs in the testing set for quantitative evaluation.

4.4 Results

In the training set, labels of dermis and epidermis are present in the ground-truth labels, which are used for the extraction of epidermis in Stage 2-Epidermis and extraction of dermis from stage 2-Epidermis. However, for the testing set, the generated segmentation mask of stage 1 must be used to extract dermis and epidermis in their corresponding stage 2 branches. Since the important tissues that we aim to segment in stage 2 are DMN in dermis and EPN in epidermis, those entities are extracted from stage 2 and are overlaid on the stage 1 segmentation mask to generate the final segmentation mask. Figure 4.1 shows the application of the trained model on the testing set. As the final post-processing step, the separate crops of the ROIs are merged back to the original shape of the ROI.

Figure 4.4 shows some examples of the original ROI in the testing set, the corresponding coarse and sparse annotations provided initially, the corresponding new full annotations

Table 4.1: Evaluation of the segmentation model on ROI testing set.

Segmentation stage	**Dice score**	**IoU**
Stage 1 (all tissues)	0.942	0.906
Stage 2-Dermis (DMN)	0.558	0.638
Stage 2-Epidermis (EPN)	0.332	0.558

(available only on DMN and EPN), and the segmentation mask generated by our model, which was trained on the coarse and sparse annotations. Quantitative results are shown in Table 4.1.

4.5 Generating WSI Segmentation Masks

The final goal of this work is to train a segmentation model on ROI images with sparse and coarse labels and produce segmentation masks for WSIs. To this end, we used the validation pipeline in Figure 4.1 on WSIs to generate a segmentation mask of Stratum Corneum, Dermis, Epidermis, Dermal Nests, and Epidermal Nests. To feed the images to the segmentation model, first, a threshold-based method was applied on each WSI to extract individual slices as explained in Section 4.5.1; then the same preprocessing as on the ROI images was applied on individual slices of the WSI. After the preprocessing, the crops were fed to the model, and after acquiring the segmentation masks, they were merged to create a WSI segmentation mask.

4.5.1 Extraction of Individual Slices

Prior to generating the segmentation masks, each whole slide biopsy image was split into individual slices using a slice extraction method. There are two benefits in performing the slide segmentation: 1) We reduced the size of the input images, 2) we can eliminate the effect of the slides' orientations since this information does not aid in the model's prediction of the

diagnosis. Figure 4.5 shows an example of a WSI containing three individual slices, which are extracted before feeding to the segmentation model.

4.5.2 Subjective Assessment with Pathologists

Since full annotations for the entire WSIs are not available for our dataset, to evaluate the WSI segmentation results qualitatively, three of our expert dermatopathologists were asked to review the segmentation masks on the WSI validation set containing 111 WSIs and grade the model's performance on several areas and tissue structures using discrete scoring. These dermatopathologists (C. May, O. Chang, S. Knezevich) are different from the original dermatopathologist (M. Mokhtari) who provided sparse annotations on the dataset and full nest annotations on a set of test ROIs. Their task was to evaluate the segmentation of the whole slide images.

To create the surveys and distribute the work, the validation set was divided into three subsets of 37 images without any overlap for each dermatopathologist to review, preserving the distribution of diagnosis class over each subset. For each dermatopathologist, an individual survey in Google Forms was provided with their corresponding subset. Each WSI was evaluated regarding four segmentation tasks: Epidermis (EP), Dermis (DE), Epidermal Nest (EPN), and Dermal Nest (DMN), chosen as being most important for diagnosis. For each segmented structure label, the dermatopathologists were asked to answer two questions with an objective of seeing if the model is over-segmenting or under-segmenting:

- **Q1:** How much of the tissue/area that is present in the corresponding WSI has been correctly identified by the model? Rate Low, Medium, or High.

- **Q2:** How much of the label identified by the model is the correct tissue/area? Rate Low, Medium, or High.

The results from these three surveys were analyzed, both individually and in combination. To translate the qualitative grading into a subjective assessment that can be used to plot

visual bar charts, we provided a numerical conversion as follows: if the grade of a label is low, the numerical equivalent is 1, medium is 2, and high is 3. The numerical equivalents of these ratings for each label in all the images were used to generate an Opinion Score (OS) which is the arithmetic mean of each label rated by the dermatopathologists (Equation 4.3), where R_n are the individual ratings for a given tissue structure, and N is the number of cases in the corresponding survey. Figure 4.6 shows the OS for each label in terms of **Q1** and **Q2** for individual pathologists and their combination.

$$OS = \frac{\sum_{n=1}^{N} R_n}{N} \tag{4.3}$$

A close examination of the WSI segmentation masks (Figure 4.7) shows that the sparse and coarse annotations provide the possibility of segmenting the tissue structures with high-quality performance on whole slide images. However, the presence of different types of noise due to coarse labeling in the training set, such as inaccurate borders and unintentional human error in labeling, plus the lack of labels on entities that are similar to dermal nests, such as inflammatory cells and eccrine ducts, result in over-labelling of nests overall. The over-labelling of the epidermal nests is higher than that of the dermal nests, which follows the pattern of our training ground-truth, in which epidermal nest annotations are noisier than dermal nest annotations. While having high sensitivity (i.e. finding all the nests) is critical in medical dataset analysis, having high specificity (i.e. reducing the false positives) is also required for accurate diagnosis. Hence, reducing noise from even sparse annotations is an important step before training a segmentation model. This can be done by having the ground truth checked by a separate pathologist from the one who created it.

4.6 Summary

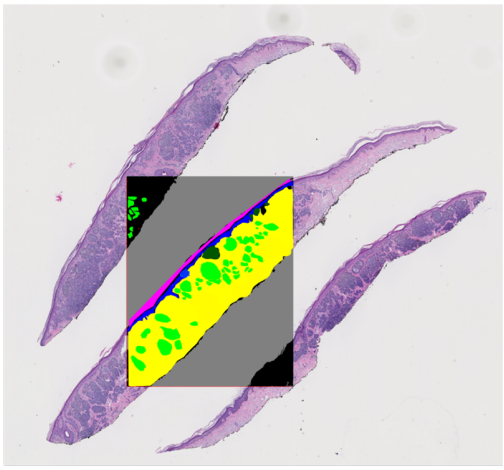
As the number of melanoma cases continues to increase, the accurate diagnosis of melanocytic lesions in skin biopsies is becoming more critical for patient care and treatment. For the pathologist, a crucial step in interpreting a melanocytic proliferation involves assessing the microanatomic location of the melanocytic population, including whether the process involves

the epidermis, dermis, or both. The semantic segmentation of these tissues (e.g. epidermis and dermis) and melanocyte position (e.g. epidermal nests and dermal nests) in skin biopsies is a required initial step in creating an automated diagnostic tool that has the potential to assist pathologists in their evaluation of melanocytic lesions, including melanoma and its precursors. Automated diagnosis tools have the potential to assist pathologists in their diagnoses.

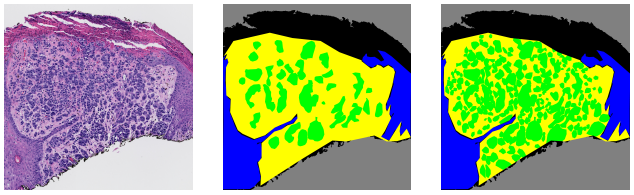
While segmentation is a significant element in the diagnosis pipeline, training a segmentation model generally requires a large, high-quality annotated ground-truth. However, most medical datasets require expert-level annotation as ground-truth, and such a requirement is a challenging, time-consuming, and expensive task, leading to a scarcity of sufficiently-sized and carefully-annotated datasets for training; overcoming this challenge is a necessity in medical image research to produce computer-aided diagnosis systems. Hence, a segmentation pipeline that can use coarse and sparse annotation to produce a segmentation model is likely to be quite beneficial.

In this work, we proposed a two-stage pipeline for the segmentation of important tissue structures in skin biopsy images using coarse and sparse annotations on small regions of WSIs. In this pipeline, larger entities were trained in the first stage, and smaller entities were trained in two sub-branches. The testing segmentation results, both on the ROIs and the WSIs, show the potential of this pipeline. Dermal Nests (DMN) and Epidermal Nests (EPN), alongside Dermis and Epidermis, are important tissues/areas in the histopathology of skin biopsy images that play a crucial role in the diagnosis. Since the ground-truth for WSIs was not available, we provided qualitative surveys for our pathologists with two questions on each tissues structure in order to evaluate the performance of the system. Question 1 (Q1) is related to recall and Question 2 (Q2) is relatead to precision. For both Dermal Nests and Epidermal Nests, the pathologists reported pretty high recall but low precision. This means that the model found most of the nests but it also found other structures that were not correct. This is partly due to inaccuracies in our training data, which are currently being corrected. Larger numbers of data will also help this problem.

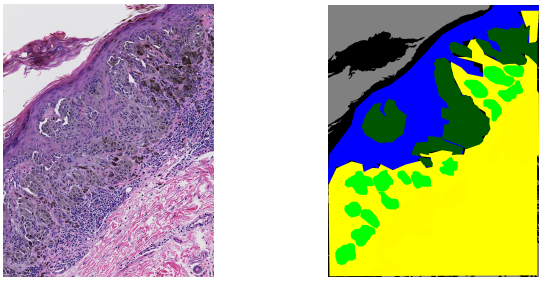
Our system was able to generate segmentation masks for both epidermis/dermis and nests with high-quality performance, indicating that having sparse annotation on important tissues has the potential for producing a useful segmentation model. On the other hand, our results suggest that both the DMN and EPN can be over-labeled by the model, highlighting the problems that coarse annotation can cause for the system, especially on a small dataset in which the ground-truth did not clearly distinguish between nests and other similar structures. These two findings suggest that having sparse, but fine, annotation on a small region of the WSI may be enough for training a better segmentation model.



(a)

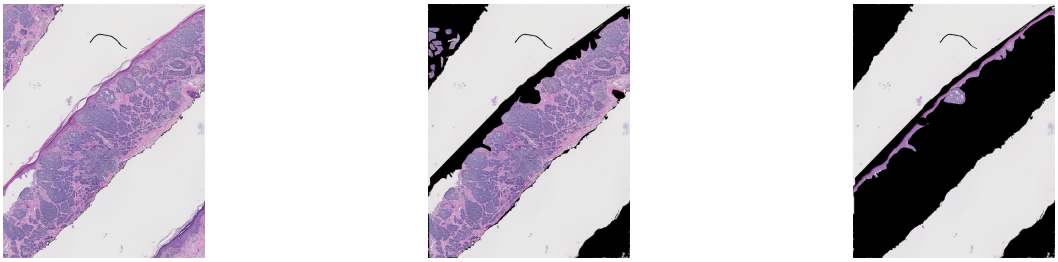


(b)

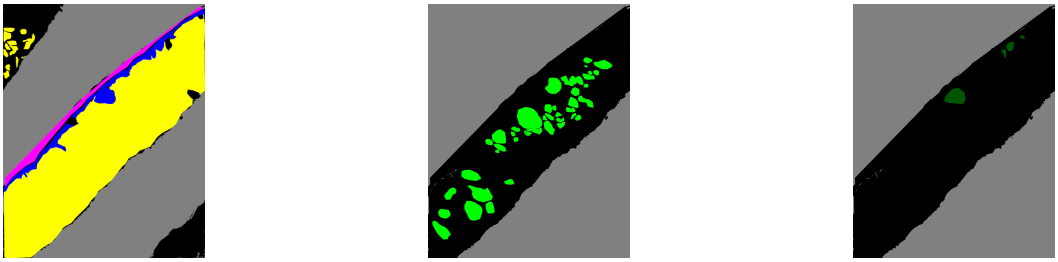


(c)

Figure 4.2: Examples of sparse and coarse annotation in our ground-truth. **(a)** Sparse annotation of an ROI overlaid on the corresponding WSI. **(b)** Example of an ROI (**left**) with its corresponding sparse annotation of Dermal Nests (DMN) (**middle**) and full annotation of Dermal Nests (DMN) (**right**); this full annotation was acquired for the sake of comparison and is not available in the training set. **(c)** An example of coarse annotations (**right**) of different tissues in an ROI (**left**).



(a) Stage 1 input image (b) Stage 2-DE input image (c) Stage 2-EP input image



(d) Stage 1 ground-truth (e) Stage 2-DE ground-truth (f) Stage 2-EP ground-truth

Figure 4.3: Examples of input images and their corresponding ground-truth for the proposed two-stage pipeline. (a) and (d) show the input image and ground-truth to stage 1, containing Dermis (DE-yellow), Epidermis (EP-blue), Corneum (COR-pink), and Background (BG-gray). (b) and (e) show the input image and ground-truth to stage 2-Dermis, containing Dermal Nests (DMN-light green), and Background (BG-gray). (c) and (f) shows the input image and ground-truth to stage 2-Epidermis, containing Epidermal Nests (EPN-dark green), and Background (BG-gray).

(a) ROI (b) Sparse (c) Fine (d) Predicted

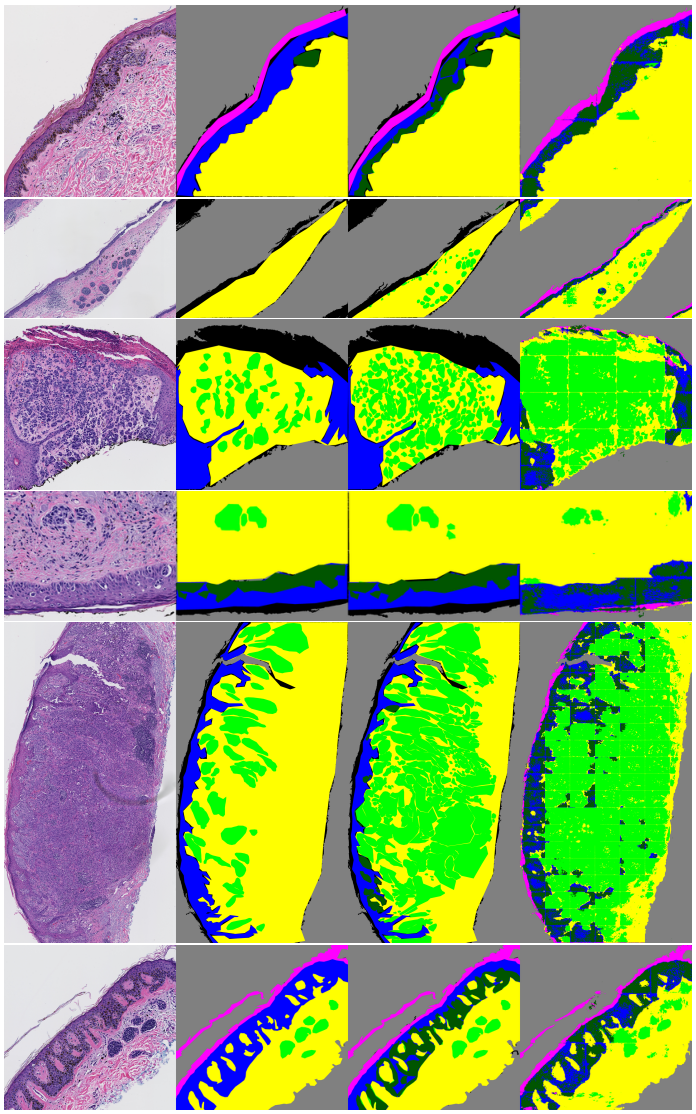


Figure 4.4: Examples of original ROI, sparse and coarse annotation, fine pixel-level nest annotation, and segmentation mask by our pipeline. The annotation and segmentation images contain Dermis (DE-yellow), Epidermis (EP-blue), Stratum Corneum (COR-pink), Background (BG-gray), Dermal Nests (DMN-light green), and Epidermal Nests (EPN-dark green). The model has been trained on sparse and coarse annotations similar to column (b) and can generate results of column (d) which are comparable to the fine pixel-level annotation of column (c). *Full annotation on nests are only available for the testing set.*

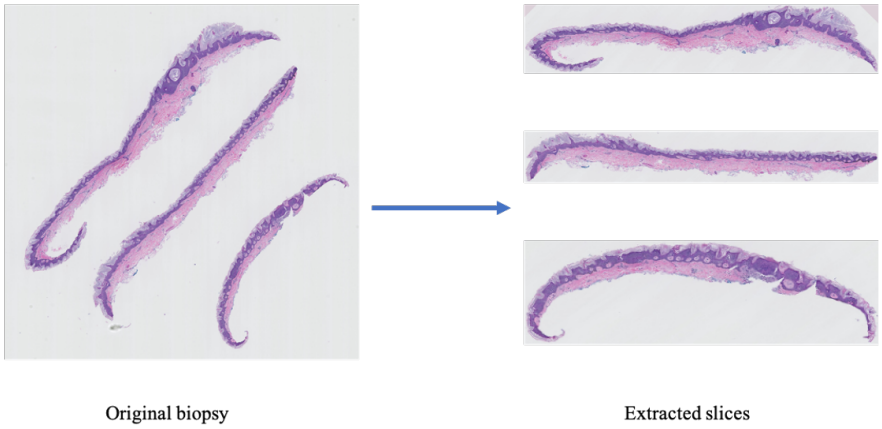
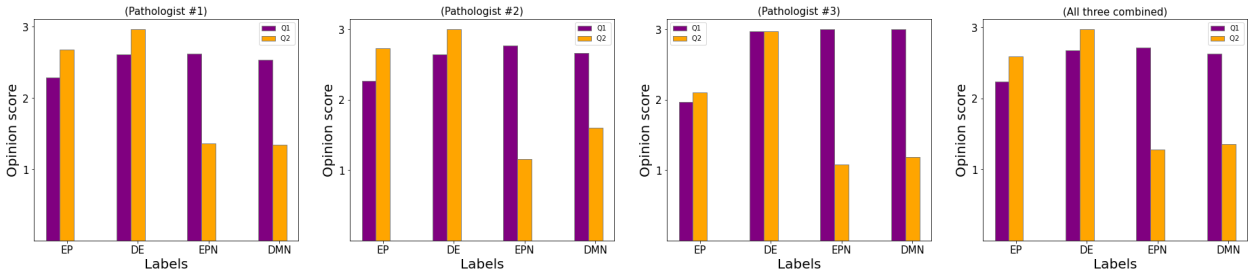
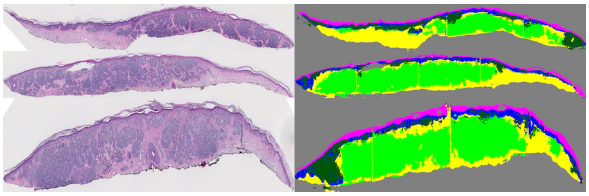


Figure 4.5: An example of a WSI (**left**) and its corresponding slice extraction (**right**).

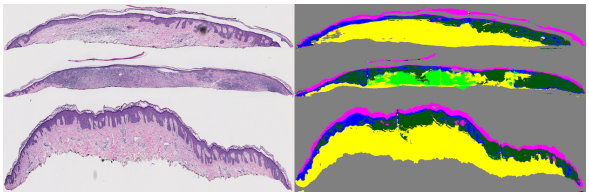


(a) Individual survey#1 (b) Individual survey#2 (c) Individual survey#3 (d) Combination

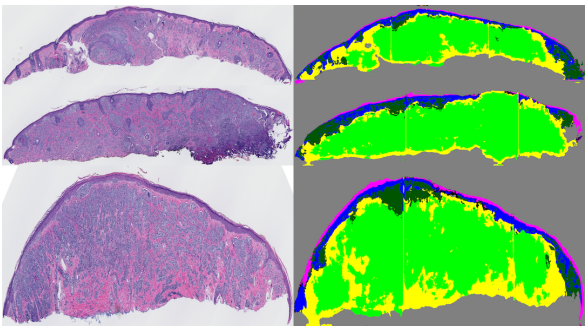
Figure 4.6: Opinion Score (OS) as subjective assessment for each label, Epidermis (EP), Dermis (DE), Dermal Nest (DMN), and Epidermal Nest (EPN), in terms of Q1 and Q2 for that tissue structure. The qualitative ratings by dermatopathologists are converted to their numerical equivalent as explained in Section 4.5.2. Each pathologist reviewed 37 different cases, (a), (b), and (c) are the individual surveys, and (d) is the combination of all three surveys.



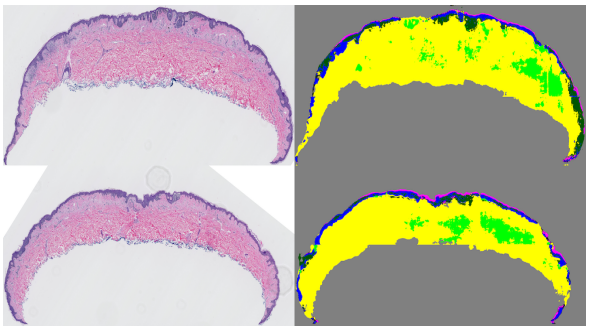
(a) DMN: High sensitivity, high specificity, EPN: High sensitivity, low specificity



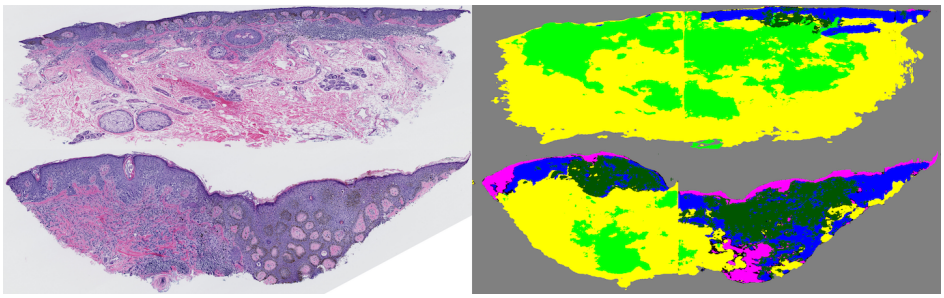
(b) DMN: High sensitivity, medium specificity, EPN: High sensitivity, low specificity



(c) DMN: High sensitivity, medium specificity, EPN: High sensitivity, low specificity



(d) DMN: High sensitivity, low specificity, EPN: High sensitivity, medium specificity



(e) DMN: Medium sensitivity, low specificity, EPN: medium sensitivity, low specificity

Figure 4.7: Examples of original WSI (left) and its corresponding segmentation mask(right). Slices of each WSI are extracted and concatenated vertically. The segmentation images contain Dermis (DE-yellow), Epidermis (EP-blue), Stratum Corneum (COR-pink), Background (BG-gray), Dermal Nests (DMN-light green), and Epidermal Nests (EPN-dark green). The model has been trained on coarse and sparse annotations. The captions show the dermatopathologists' qualitative grading on each WSI segmentation mask for dermal nests (DMN) and epidermal nests (EPN).

Chapter 5

DERMAL NEST CLASSIFICATION

5.1 Introduction

Pathologists investigate structural entities in digitized whole slide images of melanocytic skin lesions and assign a diagnosis class to the case based on the various factors, including morphological characteristics of the cells present in the biopsy images. Nests of cells, which might occur both in the dermis (dermal nests) and epidermis (epidermal nests), are one of the foremost entities in skin biopsy images as their structural features, their cell morphology, and the depth to which they have invaded the skin layers lead to a different decision on the melanoma staging.

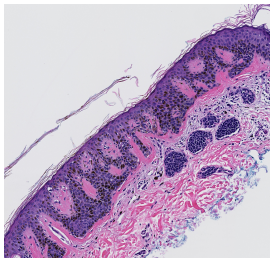
Dermal nests are one of the key components in distinguishing Mild and Moderate Nevus and Melanoma in Situ from Invasive Melanoma. Generally, dermal nests are categorized into the two sub-group of nevus nests and melanoma nests. Cases with diagnosis classes of Mild and Moderate Nevus and Melanoma in Situ only contain nevus dermal nests, while both nevus dermal Nest and melanoma dermal Nests might appear in an Invasive melanoma case.

In Chapter 4, we proposed a two-stage segmentation pipeline in which Epidermal Nests (EPN) and Dermal Nests (DMN) were segmented in its second stage. However, since not enough examples of Nevus Dermal Nests (DMN-N) were available, especially compared to other entities such as Melanoma Dermal Nests (DMN-M) and Epidermal Nests (EPN), we decided to combine Nevus Dermal Nests (DMN-N) and Melanoma Dermal Nests (DMN-M) into one class of Dermal Nests (DMN) in that project. In this chapter, we propose an additional step to the output of our segmentation model that allows us to classify segmented DMNs into two sub-categories of nevus or melanoma.

5.2 Dataset

To train a dermal nest classifier, some ground truth on different categories of dermal nests is required. The ground-truth annotations of this chapter are a subset of the coarse and sparse annotations that were introduced in section 4.2.1. The original set contained a small set of examples on nevus dermal nests on ROIs with the diagnostic class of Mild and Moderate Nevus or Melanoma in Situ, while a rather larger set of examples on melanoma dermal nests on ROIs belonged to cases with Invasive Melanoma Stage T1a and Invasive Melanoma Stage $\geq$ T1b diagnostic class. The main challenges in working with the aforementioned dermal nest's annotations were two-fold 1) the huge gap between the sample size of nevus dermal nests and melanoma dermal nests in which melanoma dermal nests contained $\sim$400M pixels which is eight times nevus dermal nests with $\sim$50M pixels. 2) No examples of nevus dermal nests on Invasive Melanoma Stage T1a and Invasive Melanoma Stage $\geq$ T1b cases were annotated. The only examples of dermal nests annotation in these classes belonged to melanoma dermal nests, while in reality, both types of nest can be present in one Invasive melanoma case. Hence, in the segmentation model of the Chapter 4, all dermal nest annotations were combined into a single class of Dermal Nests (DMN). Figure 5.1 shows example annotation of Nevus Dermal Nests (DMN-M) (Figure 5.1b) and Melanoma Dermal Nests (DMN-M) (Figure 5.1e) and their conversion to Dermal Nests (DMN) ((Figure 5.1c) and (Figure 5.1f)) which were used for the Chapter 4 dataset.

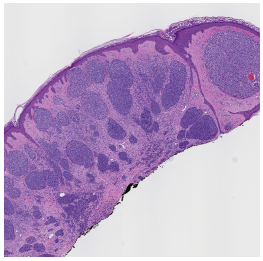
In this paper, instead of combining the two types of dermal nests, we kept them separate and extracted them into two categories of Nevus Dermal Nests (DMN-M) (Figure 5.1b) and Melanoma Dermal Nests (DMN-M) (Figure 5.1e). For the nest extraction step, after masking out everything other than dermal nests in the ROIs, we sampled the nests into two classes of "nevus" and "melanoma" using the con-nected component method. The sampling window size is 100×100. As expected, there was a noticeable imbalance in the final dataset between the two classes of "nevus" and "melanoma" nests. The number of extracted nevus nests were 604 samples while the number of extracted melanoma nests were 5732 samples. To solve this



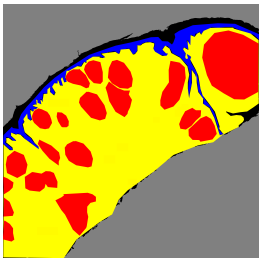
(a) Moderate Nevi ROI

(b) Nevus nest annotation

(c) Converted dermal nest



(d) Invasive $\geq$ T1b ROI



(e) Melanoma nest annotation

(f) Converted dermal nest

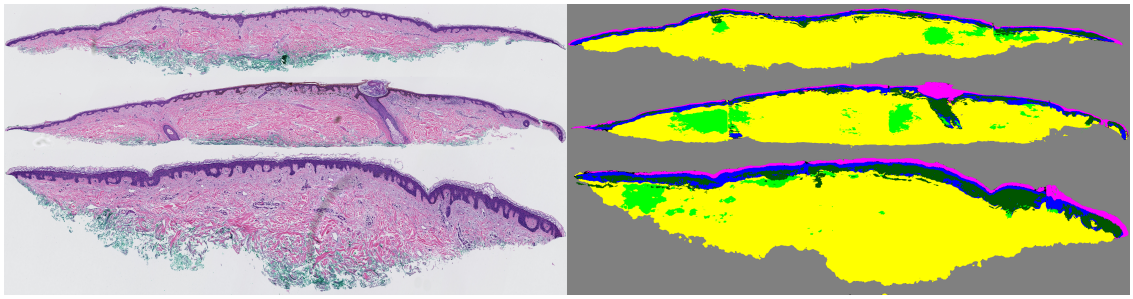
Figure 5.1: Examples of input ROI images and their corresponding annotations. (a) shows a Moderate Nevi ROI image (b) shows the original Nevus dermal nest annotation in purple, (c) is the converted version of (b) in which purple annotation of Nevus Dermal Nests(DMN-N) are converted to green markings of Dermal Nest (DMN) (d) shows an Invasive stage $\geq$ T1b ROI image (e) shows the original Melanoma nests annotation in red, (f) is the converted version of (e) in which red annotation of Melanoma Dermal Nests(DMN-M) are converted to green markings of Dermal Nest (DMN).

imbalanced dataset issue, we used the result of our previous segmentation model as explained in section 5.2.1.

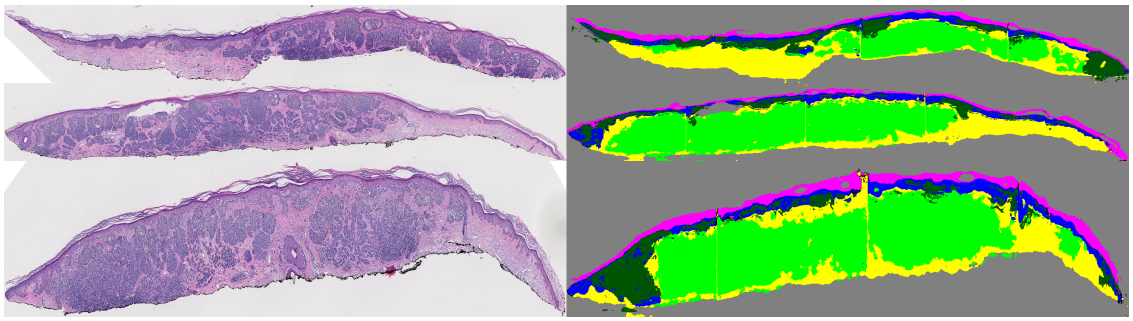
5.2.1 Solving nest sample imbalance in the training dataset

After acquiring the segmentation model output, the opportunity of overcoming the annotation imbalance in dermal nests arises. It is known that cases with a diagnosis class of Mild and Moderate Nevi or Melanoma in Situ only contain nevus dermal nests while both nevus dermal nests and melanoma dermal nests can appear in an Invasive Melanoma case. Although the segmentation model of Chapter 4 does not distinguish between nevus dermal nests and melanoma dermal nests, we know that all the nests on Mild and Moderate Nevus and Melanoma in Situ cases are nevus dermal nests. The reason is that if there is any appearance of a melanoma dermal nest on a case, that case will move to one of the invasive melanoma diagnostic categories. Figure 5.2a shows an example of segmented dermal nests on a Moderate Nevi case in which we assume all of these nests to be of nevus type based on the diagnosis of the case. Figure 5.2b shows an example of segmented dermal nests on an Invasive Melanoma Stage $\geq$ T1b case. These categories of cases are not usable for training in this project, since it is not specified which part of the segmented dermal nests are nevus and which part are melanoma.

Since the training and testing split of the dataset is consistent throughout all the projects, in addition to the fact that all the dermal nests in cases with diagnosis class of Mild and Moderate Nevus and Melanoma in Situ must be nevus dermal nests, it is only logical to apply the trained segmentation model from Chapter 4 on training WSI of Mild and Moderate Nevus or Melanoma in Situ cases, acquire Dermal Nests (DMN), extract DMNs, and re-label them as nevus dermal nests. Using the new nevus dermal nests, we can randomly extract DMN-N samples and add them to the nest classification training set to reach a balanced number of samples for both classes of DMN-N and DMN-M in the training set.



(a) Segmentation of a Moderate Nevi case



(b) Segmentation of an Invasive Melanoma case

Figure 5.2: Example of dermal nest segmentation (in light green) on WSI (a) a Moderate Nevi case; all the dermal nests are nevi type. (b) an Invasive Melanoma Stage $\geq$ T1b case; the segmented dermal nests might contain both nevi and melanoma dermal nests.

5.3 Method and Model

In our experiments, we investigate two different approaches to classify dermal nests: 1) **A feature-based** approach in which we extract features of each sample, and using conventional statistical models, we trained a classifier. 2) **A CNN-based** approach that uses well-known Convolutional Neural Networks (CNNs) as the classifier. We then evaluated each model's performance on the same testing dataset and compared the results.

5.3.1 Feature-based Classifiers

Various models in machine learning use extracted features from a dataset to learn a classification task. In this work, we utilized three different methods as follows:

- **Logistic Regression:** Logistic Regression [15] is a type of statistical analysis that is often used for predictive modeling in machine learning. This approach uses a logistic regression equation to understand the relationship between the dependent variable and one or more independent variables by estimating probabilities. We used *scikit-learn 1.0.2* [63] Python library to build a logistic regression model with L2 penalty term, saga solver, and maximum iterations of 2000.

- **Support Vector Machines (SVM):** Support vector machines (SVMs) [14] are a set of supervised learning methods used for various tasks in machine learning, such as classification, regression, etc. The objective of the SVM algorithm is to find a hyperplane in N-dimensional space (N is the number of features) that distinctly classifies the data points. For our experiments, we used *scikit-learn 1.0.2* Python library to build an SVM model using LinearSVC with L1 penalty term, squared_hinge loss function, and maximum iterations of 5000.

- **Random Forest:** Random forest [30], consists of an ensemble of a large number of individual decision trees. In this classification method, each tree in the random forest

vote for a class prediction, and the class with the most votes becomes the model's final prediction. In our experiments, we used *scikit-learn 1.0.2* Python library to build a Random Forest ensemble model with 100 decision trees, the maximum depth of the tree as None and the number of features to consider when looking for the best split to be the maximum number of features.

To extract features which will be used to train, validate, and test the aforementioned models, we used a PyTorch torchvision [47] pre-trained CNN model, ResNet18 [27] trained on ImageNet dataset [17], and extracted features. Saved features with corresponding labels of "nevus dermal nest" and "melanoma dermal nest" were used to train our classifiers.

5.3.2 CNN-based Classifiers

Another approach to training a classifier is using Convolutional Neural Networks (CNNs) since they have shown good performance in various computer vision and machine learning tasks. We trained three different architectures, using PyTorch torchvision [47] pre-trained CNN models, trained on ImageNet dataset [17]:

- **DenseNet:** densely connected convolutional neural network [31], introduces a novel connectivity mechanism to improve the flow of information between different stacked convolutional layers. In our experiments, we used a pre-trained torchvision *densenet161* architecture as a nest classifier model.

- **ShuffleNet:** ShuffleNet [97] is a convolutional neural network that utilized two new operations, point-wise group convolution and channel shuffle, to reduce computation cost while maintaining accuracy. We used a pre-trained torchvison *shufflenet_v2* for our experiments.

- **ResNet:** A residual neural network [27] is a CNN that utilizes skip connections to jump over some layers. We used a pre-trained torchvison *resnet18* for two of our experiments with different training datasets.

As the preprocessing step, we included random cropping, random rotation, horizontal flip, and normalization in the Dataloader function. All the models were trained for 20 epochs with Cross-Entropy [16] as loss function, Adam optimizer [34] with a learning rate of 0.001.

5.4 Results

All the models from both approaches were evaluated by a testing set of ROI images that was kept unseen from the model during the training process. Note that in the testing dataset, no nest samples from the segmentation model are included. The testing dataset only contains extracted nests from ROIs in which we had a pathologist's annotation as ground-truth to compare model prediction against them. Using the model with the best performance on ROI images, we generate DMN-M and DMN-N on extracted slices of the WSI.

5.4.1 Quantitative results on ROIs

All the trained models were evaluated on the same ROI testing set. Each nest classifier's performance was measured using these metrics: F-score, precision, sensitivity (recall), and specificity:

- $Sensitivity(recall) = TP/(TP + FN)$
- $Specificity = TN/(TN + FP)$
- $Precision = TP/(TP + FP)$
- $F\_score = 2 \times \frac{Precision.Recall}{Precision+Recall}$

The results of this evaluation with Statistical models and CNN models are summarized in Table 5.1 and Table 5.2, respectively.

Table 5.1: Quantitative nest classification results on ROIs-Statistical models

Approach	Method	F_score	Precision	Sensitivity	Specificity
Feature-based	**Logistic Regression**	**0.74**	**0.70**	**0.78**	**0.58**
	SVM	0.68	0.66	0.69	0.55
	Random Forest	0.71	0.65	0.78	0.45

Table 5.2: Quantitative nest classification results on ROIs-CNN models

Approach	Method	F_score	Precision	Sensitivity	Specificity
CNN-based	DenseNet	0.88	0.87	0.89	0.82
	ShuffleNet	0.78	0.80	0.76	0.74
	ResNet	**0.96**	**0.95**	**0.97**	**0.93**

5.4.2 Qualitative results on WSI

After acquiring our best nest classifier (ResNet(B)), we ran the model on all the Dermal Nests (DMN) extracted from the previous segmentation mask of Invasive Melanoma Stage T1a and $\geq$ T1b WSIs to generate Melanoma Dermal Nests (DMN-M). Any segmented DMN samples in these classes that are not classified as a DMN-M by the nest classifier model will be assigned to Nevus Dermal Nest (DMN-N). Figure 5.3 shows examples of an extracted slice of invasive melanoma WSI, corresponding Dermal Nest mask generated by our previous segmentation model, Melanoma Dermal Nest (DMN-M) portion of the Dermal Nest (DMN) as a result of nest classifier output, and Nevus Dermal Nest (DMN-N) portion of Dermal Nest (DMN) as a result of the complement of DMN-M on DMN.

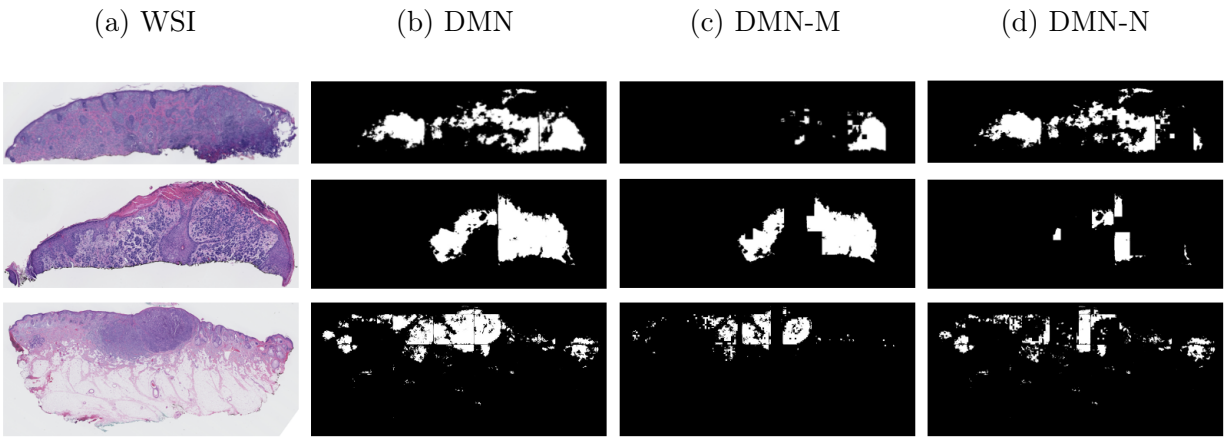


Figure 5.3: Examples of our best nest classifier, **ResNet(B)** results on WSI (a) extracted slices of invasive melanoma WSIs (b) Dermal Nest results of Chapter 4 segmentation model (c) Melanoma Dermal Nest (DMN-M) portion of DMN (d) Nevus Dermal Nest (DMN-N) portion of DMN.

5.5 Summary

Various entities play important roles in a pathologist's decision about the diagnosis of melanocytic skin lesions. Dermal nests are one of the foremost structures in biopsy images which can distinguish between a case being cancerous or non-cancerous. Furthermore, the staging of a cancerous case depends on the depth of invasion of the dermal tumor. Hence, segmentation of dermal nests, and further investigation of the dermal nest type (i.e. nevus or melanoma) is of interest.

In Chapter 4, we proposed a two-stage segmentation pipeline in which Epidermal Nests (EPN) and Dermal Nests (DMN) were segmented in its second stage. However, the necessity to further study the nature of dermal nests in different diagnosis categories led us to design a dermal nest classifier that will be able to classify segmented dermal nests (DMN) into two sub-categories of Nevus Dermal Nests (DMN-N) and Melanoma Dermal Nests (DMN-M).

Applying the trained nest classifier on WSI masks, we obtain DMN-M and DMN-N masks which have the potential to be utilized in the diagnosis pipeline. In Chapter 6 we will investigate the benefits of classified nests and possible use in improving the WSI diagnosis of skin biopsy images.

Chapter 6

IMPROVING WSI DIAGNOSIS USING TISSUE SEGMENTATION

6.1 Introduction

Deep learning and artificial intelligence have achieved unparalleled success in various tasks such as classification, segmentation, detection, etc. However, though the state-of-the-art approaches in this field show fast and accurate performance, they face challenges in dealing with medical datasets. Medical datasets usually are small in sample size, have large images, and do not have many examples of perfect annotations. However; as the field of AI in healthcare has grown significantly in recent years, more robust methods in this area have emerged.

In addition, demand for diagnosis models and classification tools based on histopathological images has increased due to the inter- and intra-variability in pathology and the potential solution that AI methods can produce. Providing prognostic and diagnostic information at the time of cancer diagnosis has important implications on patient outcomes, as automated machine learning methods on whole slide images provide a promising way forward for efficient and robust pathology analysis.

Various studies have introduced diagnosis models based on WSI and histopathology images. In [52], introduced a CNN-based deep feature extraction framework build slide-level feature representations via weighted aggregation of the patch representations and overcome the challenge of working with variable-sized regions of interest. [38] extracted relevant patch representation using self-supervised contrastive learning and introduced a dual-stream architecture with trainable distance measurement to train an MIL model called Dual-Stream Multiple Instance Learning Network (DSMIL). [9] proposed a Multiple Instance Learning

(MIL) based on transformer that first selects the top-k patches, and then uses these patches for instance-learning and bag-representation learning. In addition, this method uses a center loss that maps embeddings of instances from the same bag to a single centroid and reduces intra-class variations for final diagnosis.

Segmentation-based methods are another approach that has been studied in the field of histopathology image analysis as different tissues, and entities in these images might play an important role in the diagnosis of the case. Several works with this approach first generate semantic segmentation masks on WSI, and using the extracted information from those masks, produce an image-level diagnosis [93, 53, 57]. While this approach is a valuable study path, the challenge of dealing with imperfect annotation or lack of annotation is not addressed in such studies.

In this work, we aim to utilize some of our previous projects to improve the diagnosis of skin biopsy WSI. To this end, we incorporate the tissue segmentation masks of Chapter 4 which were generated based on sparse and coarse annotation, along with the nest classification results of Chapter 5, to investigate the potential of providing this information on WSI diagnosis of skin biopsy.

6.2 Dataset

The WSI dataset that was used in this chapter is the dataset described in Chapter 2 with 5 diagnostic classes of 1) Class I: mild dysplastic nevi, 2) class II: moderate dysplastic nevi, 3) Class III: melanoma in situ, 4) Class IV: invasive melanoma stage T1a, and 5) Class V: invasive melanoma stage $\geq$ T1b. The only difference is that since the clinical risk for progression of both Class I and Class II is extremely low, and we have a limited sample size in the aforementioned classes, we regrouped the five classes to four diagnostic classes by combining samples from class I and II into one class. The final 4 classes will be 1) Class I-II: mild and moderate dysplastic nevi (MMD), 2) Class III: melanoma in situ (MIS), 3) Class IV: invasive melanoma stage T1a (T1a), and 4) Class V: invasive melanoma stage $\geq$ T1b (T1b).

As mentioned in section 4.5.1, to 1) reduce the size of the input images, and 2) eliminate

the effect of the slides' orientations, since this information does not aid in the model's prediction of the diagnosis, we used extracted slices from WSIs. An example of a WSI and its corresponding extracted slices is shown in Figure 4.5.

The main resolution which we used to extract individual slices was 20x. Using this resolution, we extracted lower resolutions of 7.5x, 10x, and 12.5x which we later used for our experimental studies.

6.2.1 Binarized Segmentation Masks

The segmentation masks generated by the proposed pipeline in Chapter 4 were used in the current project. Each tissue mask from that project (Epidermis (EP), Dermis (DE), Epidermal Nest (EPN), and Dermal Nest (DMN)) was separated into a single binary mask in order to have more control over tissue combination in our experimental studies on the diagnosis accuracy. In addition to the aforementioned tissue masks, we included the two types of dermal masks from Chapter 5 as two separate binary masks of Nevus Dermal Nest (DMN-N) and Melanoma Dermal Nest (DMN-M). Figure 6.1 shows examples of binary masks for two classes of Mild and Moderate Nevi (MMD) and Invasive Melanoma Stage $\geq$ T1b (T1b). Note that the MMD case does not include any DMN-M; hence, the corresponding mask is all zeros.

6.2.2 Dataset Split

The dataset of WSIs before extraction of slices are divided into half, conserving the original set's diagnostic class distribution over both subset. One-half of the dataset is used for training and validation subsets, and the other half of the dataset is kept unseen from the model during the training and solely used for the final evaluation of the trained model. This split is kept fixed over all the experiments. After splitting the dataset, the extraction step that is explained in section 4.5.1 is applied over all the WSIs in the training, validation, and testing subsets.

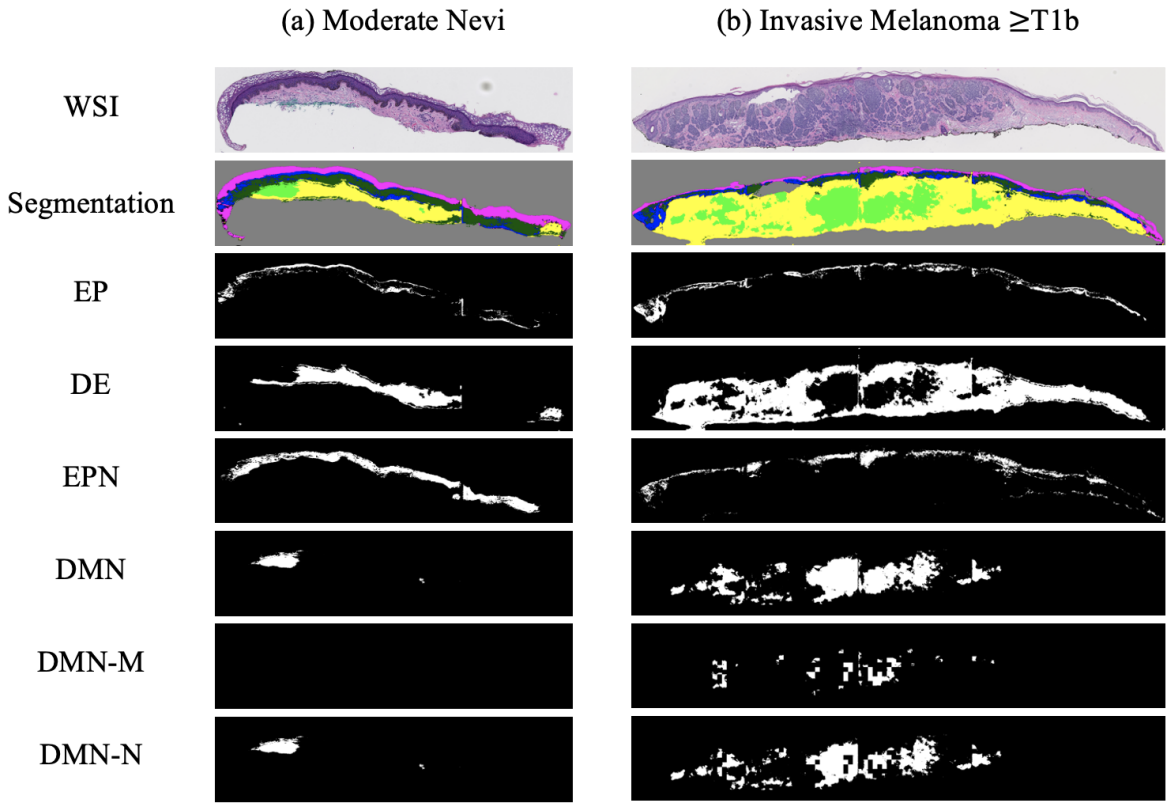


Figure 6.1: Examples of binarized segmentation masks (a) a Moderate Nevi case (b) an Invasive Melanoma Stage $\geq$ T1b. From top to bottom, one extracted slice from a WSI, all segmentation masks in one mask (containing EP, DE, EPN, and DMN), binary Epidermis (EP) mask, binary Dermis (DE) mask, binary Epidermal Nest (EPN) mask, binary Dermal Nest (DMN) mask, binary Melanoma Dermal Nest (DMN-M), and binary Nevus Dermal Nest (DMN-N) mask are shown.

6.2.3 Soft Labels

Usually, each WSI has multiple slices; however, not all the slices contain related information to the assigned diagnostic class to the case. In our dataset, the ROIs (some examples in Figure 2.2) that helped pathologists in diagnosis belong to one or two tissue slices, while the other tissue slices may correspond to other diagnostic categories. If all the extracted slices from a WSI are assigned to one diagnostic class, there is the risk of false representation of that diagnostic class which can interrupt the learning process of a model. To handle the aforementioned issue, we used a method that was previously developed by our group in which using a singular-value decomposition (SVD), soft labels will be assigned to the slices that do not have an ROI on them. For more information about the details of this method, refer to [91].

6.2.4 Combining WSI and Segmentation masks

We tried various methods to combine the information from WSI and corresponding segmentation masks' information. The final method that we chose to implement and run our experiment is as follows: Each WSI has 3 channels of RGB: Red (R), Green (G), and Blue (B). In order to add segmentation mask information to our data, we concatenate each mask as a new channel to the image. For example, if we add a DMN channel to the WSI, we will have a new input with 4 channels: R, G, B, and DMN. This approach gives the flexibility of investigating any combination of tissue masks that is of interest. In addition, the feature extractor obtains the information of appended tissue masks along with the original WSI which might result in a more representative feature set.

6.2.5 Feature extraction

We use MobileNetv2 [74] pre-trained on the ImageNet dataset [17] as a feature extractor on our extracted patches. MobileNetv2 outputs 1280-dimensional patch-wise features after global average pooling. Since the pre-trained network on the ImageNet dataset is essentially

a network with 3 input channels of RGB, we modified the first layer of the network by replacing it with a *Conv2d* layer that has input channels equal to the number of input image channels. The number is not fixed since as explained in section 6.2.4, the number of input image channels depends on the tissue mask combination in a specific experiment. Changing the first layer of the network which is not pre-trained on any image has the potential of negatively impacting the feature extraction step; however, as we will see in the next sections, the results do not show any clear effect of such. The reason might be the nature of CNNs in which the first few layers are focused on low-level features while the middle layers mainly extract high-level and fine detailed features.

6.3 Method and Model

In this work, we aim to investigate the potential of improving WSI analysis using previously generated segmentation masks from sparse and coarse annotation. To this end, we designed several experiments with various tissue combinations and trained multiple diagnosis models using Scale-Aware Transformer Network (ScATNet) [91] as a base model.

6.3.1 Scale-Aware Transformer Network (ScATNet)

In previous work, [91] proposed Scale-Aware Transformer Network (ScATNet) for Diagnosing Melanocytic Lesions using WSI. ScATNet is designed to allow the model to learn local and global representations from multiple input scales. This end-to-end pipeline has three main steps: (1) learn local patch-wise embeddings using a CNN for each input scale, (2) learn contextualized patch embeddings for each input scale using transformers, and (3) learn scale-aware embeddings across multiple input scales using transformers [91].

ScATNet projects extracted patch-wise features explained in section 6.2.5 linearly to a 128-dimensional space and then learns contextualized patch-wise and scale-wise embeddings using transformers. For learning contextualized patch-wise and scale-wise representations, a stack of two transformer units is used. Also, in each transformer unit, the number of heads in the self-attention layer is set to 4, and the feed-forward network dimension is set to 512.

In our work, using segmentation masks along corresponding WSI, we achieved the best results using a single scale version of ScATNet. In section 6.4.1, the comparison of different tissue types and various scales will be presented.

6.3.2 Experimental studies

In order to investigate the impact of different tissue types, we designed several experiments with various combinations of tissue segmentation masks, using ScATNet as the basic model. In each experiment, we included specific segmentation masks along with the WSI, extracted the features as explained in section 6.2.5 and using the extracted features, we trained and tested a diagnosis model. We ran the experiments with various resolution scales (7.5 x, 10x, 12.5x, combination of two scales, and all three scales), with different hyperparameters, and after finding the best setting, we ran all the experiments with different random seeds.

Figure 6.2 shows an overview of our approach.

6.3.3 Hyperparameters

ScATNet was trained for 200 epochs in an end-to-end fashion using the ADAM optimizer with a linear learning rate warm-up strategy and step learning rate decay. The best result in our experimental studies was achieved using a single scale of 7.5x.

6.4 Results

As a model selection step after training each experiment for 200 epochs, and to improve the models robustness against stochastic noise, we averaged the best 5 model checkpoints within a single training process inspired by [8]. Then we evaluated all of our experiments over the same testing set. A WSI might contain multiple tissue slices, which were extracted into single slices, and each of these slices might have a different diagnostic class prediction. To decide on the final diagnosis of a specific WSI, we used max-voting, which means if one of the tissue slices in a WSI is invasive melanoma, then the entire WSI corresponds to invasive melanoma

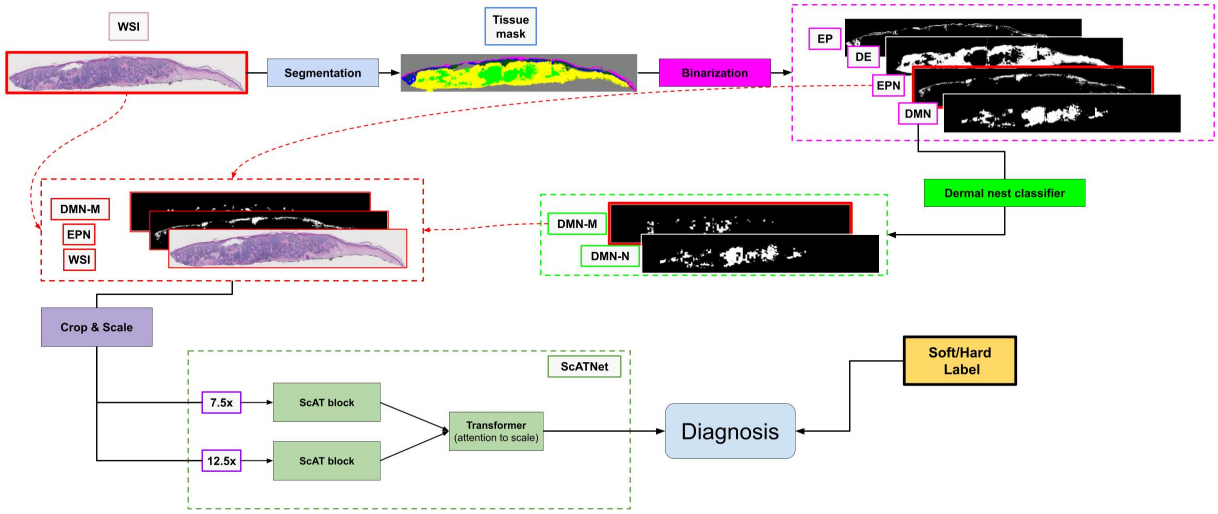


Figure 6.2: Overview of the diagnosis pipeline. The WSI goes to the segmentation pipeline to generate a tissue segmentation mask. Then, four clinically important tissue structures, Epidermis (EP), Dermis (DE), Epidermal Nest (EPN), and Dermal Nest (DMN) will be extracted into four corresponding binary masks. Extracted Dermal Nests will go through a dermal nests classification step to generate two sub-category of Melanoma Dermal Nest (DMN-M) and Nevus Dermal Nest (DMN-N). Then, the selected tissue masks based on the experiment will be concatenated along the RGB channels of the WSI image. Each image will be cropped into smaller patches afterward. The patches go through ScATNet pipeline that extracts patch embeddings, then, using contextualized patch-embedding and scale-aware embedding across available scales, predicts the diagnostic class of each individual input slice. Then using the soft labeling, the final diagnosis of the case will be chosen from Mild and Moderate Dysplastic Nevi (MMD), Melanoma in Situ (MIS), Invasive Melanoma T1a (T1a) and Melanoma Invasive $\geq$ T1b (T1b). Note that concatenated masks to the WSI (DMN-M and EPN) and ScATNet scales (7.5x and 12.5x) shown in this figure are one example our multiple experimental studies.

and cannot be MMD or MIS. This approach was inspired by how pathologists make their diagnosis decision on skin biopsy images.

6.4.1 Experimental Results

We evaluated all the models based on micro F_score, Sensitivity (recall), and Specificity. Note that in dealing with a multi-class classification, where every test datum should belong to only 1 class and not multi-label, we cannot use the same F_score as in binary class classification (i.e. macro F_score in multi-class classification). The correct way to report an F_score in multi-class classification is to calculate the micro-averaged F_score (AKA micro F_score) based on micro-precision and micro-recall. Micro-precision measures the precision of the aggregated contributions of all classes, and micro-recall measures the recall of the aggregated contributions of all classes.

- $micro\ Precision = \frac{TP_{sum}}{TP_{sum}+FP_{sum}}$
- $micro\ Recall = \frac{TP_{sum}}{TP_{sum}+FN_{sum}}$
- $micro\ F\_score = 2 \times \frac{micro\_precision \times micro\_recall}{micro\_precision + micro\_recall}$
- $Sensitivity(recall) = \frac{TP_{sum}}{TP_{sum}+FN_{sum}}$
- $Specificity = \frac{TN_{sum}}{TN_{sum}+FP_{sum}}$

The summary of results is shown in Table 6.1. F_score of each experiment is reported based on 10 different random seeds, along with average Sensitivity and Specificity over 10 random seeds per experiment. In our experiments, the (Average, Max) F_scores were (0.54,.058) for the raw WSI with no segmentation masks, which improved to a high of (0.60,0.62) for the raw WSI plus the epidermis mask and the dermal melanoma mask (i.e. the cancerous nests in the dermis). The addition of the dermal melanoma mask was important as it gave a significant gain over just providing dermal nests. Note that we started with a

rather low F_score for the raw WSI and fixed those parameters to achieve stability, so it is possible that even higher values can be achieved in the top row by starting with a different set of parameters for the WSI run. However, we favored stability, and the (0.54, 0.58) scores were stable, in that they could be achieved repeatably.50.

Table 6.1: Experimental results of WSI diagnosis along segmentation masks

Experiment	**F_score**				**Sensitivity**	**Specificity**
	Average	Min	Max	Median		
WSI + EPN + DMN-M	**0.60**	**0.58**	**0.62**	**0.59**	**0.60**	**0.87**
WSI + EPN + DMN	0.57	0.54	0.61	0.56	0.57	0.85
WSI + EPN + DMN-M + DMN-N	0.56	0.53	0.60	0.55	0.56	0.85
WSI + EP + DE + EPN + DMN	0.55	0.53	0.59	0.54	0.55	0.85
WSI	*0.54*	*0.53*	*0.58*	*0.54*	*0.54*	*0.85*
WSI + EPN	0.54	0.52	0.58	0.53	0.54	0.85
WSI + DMN	0.54	0.51	0.56	0.54	0.54	0.85
WSI + DMN-M + DMN-N	0.54	0.52	0.55	0.54	0.54	0.86
WSI + DMN-M	0.52	0.50	0.55	0.51	0.52	0.84

*F_score are reported for 10 random seeds.

**Sensitivity and Specificity are average scores over 10 random seeds per experiment.

6.4.2 Comparison of Confusion Matrices

Table 6.2 shows a comparison of two experiments' confusion matrices. Table 6.2a is an example of a multi-class confusion matrix of experiments that only contain RGB channels of the WSI in the dataset, while Table 6.2b shows an example of an experiment in which we had R, G, and B channel of the WSI along with two extra channels of Epidermal Nest (EPN)

binary segmentation mask and Melanoma Dermal Nest (DMN-M) binary segmentation mask (a total of 5 channels per image).

As shown in the tables, the number of True Positives (TP) of classes MIS, T1a, and T1b increased in the experiment in which we included segmentation masks along with WSI. Another important finding is that the misclassified cases of MIS when we have EPN and DMN-M information are mostly on T1b and less on MMD. In the real world, MIS is a challenging case for pathologists to make a definite diagnosis. The comparison of confusion matrices in Table 6.2 and tissue experiments' result of Table 6.1 shows that the model is able to learn more information when segmentation masks are introduced along with the WSI, which can be an assistance to pathologists on challenging cases.

	MMD	**MIS**	**T1a**	**T1b**
MMD	**17**	8	4	0
MIS	7	**12**	9	2
T1a	0	9	**18**	4
T1b	0	2	9	**12**

(a) WSI

	MMD	**MIS**	**T1a**	**T1b**
MMD	**17**	9	3	0
MIS	3	**16**	10	1
T1a	5	2	**18**	4
T1b	0	0	8	**15**

(b) WSI + EPN + DMN-M

Table 6.2: Comparison of two confusion matrices. Rows are defined by expert consensus and columns are by model predictions. (a) an example experiment with only WSI and no segmentation mask. (b) an example experiment of WSI + EPN + DMN-M.

6.4.3 Single-Scale vs. Multi-scale

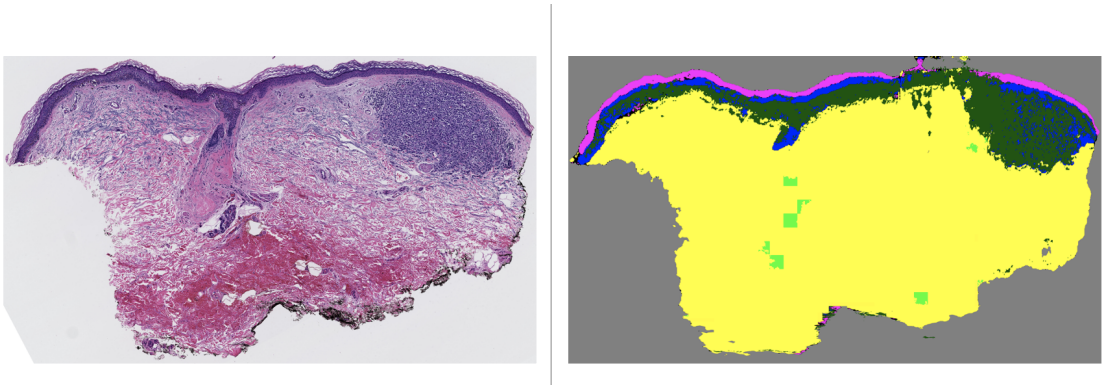
In our experiments, we ran each setting of tissue experiments with single-scale, 2 scales, and 3 scales. A summary of results for one example tissue experiment (WSI + EPN + DMN-M) in comparison with a raw WSI, which has the exact same parameters and scales, are summarized

in Table 6.3. These results suggest that having segmentation masks does not improve the performance when ScATNet is trained on multiple scales, and the gain of improvement is lower when the higher resolution of WSI along with segmentation masks is used.

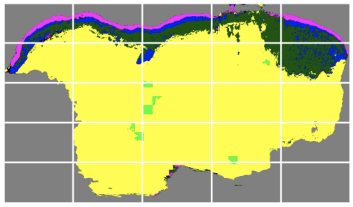
This behavior can be explained by the specific strategy of ScATNet in patching input images on different scales. For example, images in 7.5x resolution are divided into $5 \times 5 = 25$ crops while 12.5x images are divided into $9 \times 9 = 81$ crops. In addition, the transformer unit in the ScATNet architecture includes a self-attention module that learns to pay more attention (i.e. assign higher weight) to specific patches in an image. When we introduce a WSI along with its corresponding dermal nests and epidermal nests, the model learns during the training process that these structures are important in decision making. Hence, when these tissue structures appear in a testing case's segmentation mask, the model assigns higher weights to the patches that contain those structures. If a segmentation mask of a testing case is inaccurate, especially when some important structures are over-labeled, it can negatively impact the model's decision-making and lead to a false prediction. The possibility of such an impact could be higher in higher resolutions, since there will be more patches with inaccurate tissue labels; hence, higher weights on irrelevant patches. Figure 6.3 shows an example of a test set WSI and corresponding segmentation mask (Figure 6.3a) that includes dermis, epidermis, melanoma dermal nest, epidermal nest, corneum, and background. The segmentation of epidermal nest is inaccurate and over-labeled, and potentially led to a wrong prediction on resolution 12.5x (Figure 6.3c), since the number of patches with noise at that resolution is more than at resolution 7.5x (Figure 6.3b).

6.4.4 Comparison to US Pathologists

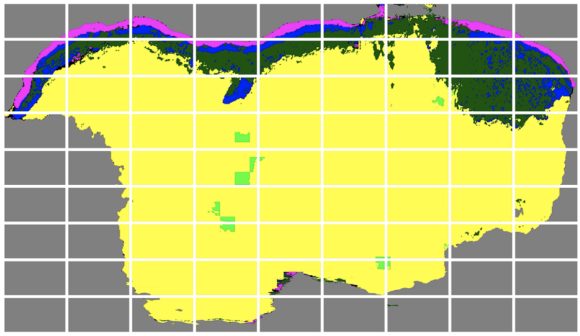
We have access to the interpretation of 187 US pathologists on the same testing set that we used in our experimental studies. Table 6.4 shows the comparison of F_score, sensitivity and specificity of pathologists' performance and our best model (WSI + EPN + DMN-M) performance. We observe that our model either outperforms the pathologists' results on the challenging classes of MIS and T1a or has a comparable performance. This finding shows



(a) a WSI and corresponding segmentation mask



(b) 7.5x scale divided into 25 crops



(c) 12.5x scale divided into 81 crops

Figure 6.3: low-resolution vs. high-resolution patching when there is an inaccurate segmentation mask is in testing case. a) a WSI and corresponding segmentation mask that includes dermis, epidermis, melanoma dermal nest, epidermal nest, corneum, and background. In this example case, epidermal nests are inaccurately segmented and over-labeled. b) the segmentation mask in 7.5x scale divided into 25 crops as input patches for ScATNet. c) the segmentation mask in 12.5x scale divided into 81 crops as input patches for ScATNet. There are higher number of patches with inaccurate and noisy segmentation on 12.5x scale compared to 7.5 scale which possibly led to a false prediction on 12.5x scale using ScATNet.

Table 6.3: Comparison of F_score results of raw WSI and tissue experiment (WSI + EPN + DMN-M) on single-scale experiments and multi-scale experiments.

Scale	**raw WSI**	**WSI + EPN + DMN-M**
7.5x	0.54	0.60
12.5x	0.56	0.57
7.5x & 12.5x	0.57	0.56
7.5x & 10x & 12.5x	0.57	0.55

the potential that providing an assistant tool can have in the time of cancer diagnosis and treatment.

Table 6.4: Comparison of F_score, sensitivity and specificity of 187 US pathologists and our best model (WSI + EPN + DMN-M) on the same testing set.

Class	**F_score**		**Sensitivity**		**Specificity**	
	pathologists	Ours	pathologists	Ours	pathologists	Ours
MMD	0.71	0.67	0.92	0.76	0.76	0.81
MIS	0.49	0.50	0.46	0.44	0.85	0.89
T1a	0.62	0.57	0.51	0.64	.95	0.79
T1b	0.72	0.67	0.78	0.57	0.97	0.96

6.4.5 Comparison to Other Baselines

We compared our results with several other methods developed to make a diagnosis based on histopathology images.

- **Weighted Feature Aggregation:** Deep Feature Representations for Variable-Sized Regions of Interest was introduced by [52]. In this method, a CNN-based deep feature extraction framework builds slide-level feature representations via weighted aggregation of the patch representations. In this pipeline, the patch-wise feature will be extracted by a VGG16 pre-trained CNN, then using two different approaches of either penultimate layer features (penultimate-weighted) or hypercolumn features (hypercolumn-weighted), the features will be concatenated in a weighted manner. As the last step, using average pooling, a slide-level representation will be generated which will later be used for training and testing the diagnosis CNN model.

- **Dual-stream Multiple Instance Learning Network (DSMIL):** In this work, [38] used self-supervised contrastive learning to extract good representations from patches and using an aggregator that models the relations of the instances in a dual-stream architecture with trainable distance measurement, trains a MIL model.

- **Multiple Instance Learning with Center Embeddings (ChikonMIL):** [9] proposed a Multiple Instance Learning (MIL) method that first selects the top-k patches, and then uses these patches for instance-learning and bag-representation learning. In addition, this method uses a center loss that maps embeddings of instances from the same bag to a single centroid and reduces intra-class variations for the final diagnosis.

The results of all the baseline methods and their comparison with our best model are summarized in Table 6.5. Our model using the epidermal nests and dermal melanoma nests is able to beat all of them.

6.5 Summary

The rapidly growing number of melanoma cases along with inter- and intra-variability of diagnosis by human pathologists is of concern in this field. On the other hand, advances in machine learning and artificial intelligence methods have presented the potential to provide

Table 6.5: Comparison of baseline methods with our best model (WSI + EPN + DMN-M)

Method	**F_score**	**Sensitivity**	**Specificity**
penultimate-weighted [52]	0.44	0.44	0.81
hypercolumn-weighted [52]	0.43	0.43	0.81
DSMIL [38]	0.50	0.50	0.83
ChikonMIL [9]	0.56	0.56	0.85
Ours	**0.60**	**0.60**	**0.87**

assistant tools for the pathologists to analyze whole slide images (WSIs) for diagnosis and prognosis objectives.

In recent years, deep learning methods have proved to have excellent performance in different tasks such as image classification. However, most of the state-of-the-art methods either require a fairly large dataset to train a model or a large amount of pixel-level annotation. Both of these requirements are a challenge in dealing with medical datasets as these datasets are usually small, especially compared to general datasets such as ImageNet [17], and obtaining fine manual annotation on them is not a time or cost-effective task.

In this work, we proposed an approach that uses the segmentation masks which we previously obtained using sparse and coarse annotation in Chapter 4, and adds information to WSI from a fairly small dataset of skin biopsy images. The goal was to investigate the potential of each important tissue mask in skin biopsy images to improve the results of a multi-class diagnosis model.

Our experiments showed that including certain segmentation masks along with WSIs yields a better diagnosis output with one scale. One of the foremost tissue types in skin biopsy images are nests which contain various types such as Epidermal Nests (EPN), Nevus Dermal Nests (DMN-N), and Melanoma Dermal Nest (DMN-M). We observed significant

improvement when including EPN and DMN-M (which is considered the cancerous type of dermal nests) along with corresponding WSI, compared to the experiments that do not include any segmentation masks. Further analysis showed that including the aforementioned entities improved the learning of the model on invasive melanoma and melanoma in situ. Melanoma in Situ (MIS) and Invasive Melanoma (T1a) are challenging classes for pathologists to make a consensus decision on those classes. Improvement in the challenging classes proves the potential AI has in healthcare and pathology.

The dataset that we used in this project is in melanocytes skin lesions and while the cases included were carefully selected to represent the full spectrum of cases in clinical practices in the US, we are not certain how this would work when the full spectrum of skin biopsy (e.g., including non-melanocytic lesions) would work. In addition, the sparse annotations for the segmentation project were provided on ROIs on the WSI, which means there was prior knowledge on which part of the WSI contains valuable information. Not all medical datasets benefit from having ROI assigned to each case.

Chapter 7

CONCLUSION

In this work, we aimed to provide prognostic and diagnostic information that has important implications in the time of cancer monitoring and treatment and can led to improving the whole melanoma diagnosis pipeline and patient outcome. We developed novel methods toward detection of cellular entities, segmentation of structural entities, and classification of whole slide images. We used 240 skin biopsy WSIs from M-Path dataset with five diagnostic categories: Mild Dysplastic Nevi, Moderate Dysplastic Nevi, Melanoma in Situ, invasive melanoma stage T1a, and invasive melanoma stage $\geq$ T1b. This dataset include one region of interest (ROI) per image which was assigned by expert pathologists when making the final diagnosis on the case.

One of the main challenges we had to tackle in this work was working with a very small dataset and limited annotation and ground-truth. In each chapter, we tried a specific approach to overcome this obstacle and showed the potential of working with such a limitation.

We first studied the classification of mitosis which is one of most important cellular entities and considered to be a key prognostic factor during assessment of a case by pathologists (Chapter 3). Using two state-of-the-art CNN-based models, ESPNet and DenseNet, we generate two separate models for mitosis classification and evaluated their performance in terms of sensitivity, specificity, and F-score. The ESPNet and DenseNet results on our melanoma dataset had a sensitivity of 0.976 and 0.968, and a specificity of 0.987 and 0.995, respectively, with F-scores of .968 and .976, respectively [58].

In Chapter 4, we developed a two-stage pipeline which utilized the newly obtained coarse and sparse annotations on ROI portion of our dataset and trains a U-Net based segmentation model. The main structural entities which were used in this work were epidermis, dermis,

epidermal nests and dermal nests, with the goal of using the results from this work in automated diagnosis systems or serve them as a diagnostic aid in the decision-making process. Applying the trained model on ROI, we generated segmentation masks on all the WSI in our dataset [59]. Later, in Chapter 5, using a CNN model (ResNet), we classified the dermal nests into two sub-category of melanoma dermal nest (AKA cancerous nest) and nevus nest (AKA non-cancerous nest) to study the potential of having these two dermal nests type in our automated diagnosis pipeline.

Finally, in Chapter 6, we utilized WSI segmentation masks from Chapter 4 and refined versions of Chapter 6 and studied the impact of adding each tissue mask to WSI in the classification of our dataset into diagnostic categories. To train diagnostic models, we used ScATNet which is a scale-aware transformer-based model developed for histopathology image diagnosis. Our experiments showed that including certain segmentation masks such as epidermal nests and dermal nests, specifically melanoma dermal nests, along with WSIs result in a better diagnosis performance.

For future work, utilizing other cellular entities such as Melanocyte cells and nests seems like a promising path in improving diagnostic accuracy. In addition, our finding showed that noisy and coarse segmentation on a small dataset like ours can negatively impact the diagnosis accuracy. Therefore, working on refining the segmentation masks or the approach of introducing them to the diagnosis pipeline might be a good strategy in improving the accuracy of model prediction.

BIBLIOGRAPHY

[1] Salah Alheejawi, Richard Berendt, Naresh Jha, Santi P Maity, and Mrinal Mandal. Automated proliferation index calculation for skin melanoma biopsy images using machine learning. *Computerized Medical Imaging and Graphics*, 89:101893, 2021.

[2] Salah Alheejawi, Hongming Xu, Richard Berendt, Naresh Jha, and Mrinal Mandal. Novel lymph node segmentation and proliferation index measurement for skin melanoma biopsy images. *Computerized Medical Imaging and Graphics*, 73:19–29, 2019.

[3] John-Melle Bokhorst, Hans Pinckaers, Peter van Zwam, Iris Nagtegaal, Jeroen van der Laak, and Francesco Ciompi. Learning from sparsely annotated data for semantic segmentation in histopathology images. In *International Conference on Medical Imaging with Deep Learning–Full Paper Track*, 2018.

[4] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and flexible image augmentations. *Information*, 11(2), 2020.

[5] Patricia A Carney, Lisa M Reisch, Michael W Piepkorn, Raymond L Barnhill, David E Elder, Stevan Knezevich, Berta M Geller, Gary Longton, and Joann G Elmore. Achieving consensus for the histopathologic diagnosis of melanocytic lesions: use of the modified delphi method. *Journal of cutaneous pathology*, 43(10):830–837, 2016.

[6] Norman Casagrande. Automatic music classification using boosting algorithms and auditory features. 2006.

[7] Hao Chen, Qi Dou, Xi Wang, Jing Qin, and Pheng Ann Heng. Mitosis detection

in breast cancer histology images via deep cascaded networks. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[8] Hugh Chen, Scott Lundberg, and Su-In Lee. Checkpoint ensembles: Ensemble methods from a single training process. *arXiv preprint arXiv:1710.03282*, 2017.

[9] Philip Chikontwe, Meejeong Kim, Soo Jeong Nam, Heounjeong Go, and Sang Hyun Park. Multiple instance learning with center embeddings for histopathology classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 519–528. Springer, 2020.

[10] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

[11] 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.

[12] Stanford CS class CS231n. CS231n Convolutional Neural Networks for Visual Recognition. accessed 2018-05-25.

[13] Germán Corredor, Xiangxue Wang, Cheng Lu, Vamsidhar Velcheti, Eduardo Romero, and Anant Madabhushi. A watershed and feature-based approach for automated detection of lymphocytes on lung cancer images. In *Medical Imaging 2018: Digital Pathology*, volume 10581, page 105810R. International Society for Optics and Photonics, 2018.

[14] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[15] David R Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232, 1958.

[16] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Annals of operations research*, 134(1):19–67, 2005.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[18] Veena Dodballapur, Yang Song, Heng Huang, Mei Chen, Wojciech Chrzanowski, and Weidong Cai. Mask-driven mitosis detection in histopathology images. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1855–1859. IEEE, 2019.

[19] Joann G Elmore, Raymond L Barnhill, David E Elder, Gary M Longton, Margaret S Pepe, Lisa M Reisch, Patricia A Carney, Linda J Titus, Heidi D Nelson, Tracy Onega, et al. Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study. *Bmj*, 357, 2017.

[20] 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.

[21] 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.

[22] Maayan Frid-Adar, Avi Ben-Cohen, Rula Amer, and Hayit Greenspan. Improving the segmentation of anatomical structures in chest radiographs using u-net with an imagenet

pre-trained encoder. In *Image Analysis for Moving Organ, Breast, and Thoracic Images*, pages 159–168. Springer, 2018.

[23] Yu Gordienko, Peng Gang, Jiang Hui, Wei Zeng, Yu Kochura, Oleg Alienin, Oleksandr Rokovyi, and Sergii Stirenko. Deep learning with lung segmentation and bone shadow exclusion techniques for chest x-ray analysis of lung cancer. In *International Conference on Computer Science, Engineering and Education Applications*, pages 638–647. Springer, 2018.

[24] 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.

[25] Gery P Guy Jr, Cheryll C Thomas, Trevor Thompson, Meg Watson, Greta M Massetti, and Lisa C Richardson. Vital signs: melanoma incidence and mortality trends and projections—united states, 1982–2030. *MMWR. Morbidity and mortality weekly report*, 64(21):591, 2015.

[26] Adel Hafiane, Filiz Bunyak, and Kannappan Palaniappan. Fuzzy clustering and active contours for histopathology image segmentation and nuclei detection. In *Advanced concepts for intelligent vision systems*, pages 903–914. Springer, 2008.

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

[28] 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.

[29] Achim Hekler, Jochen Sven Utikal, Alexander H Enk, Carola Berking, Joachim Klode, Dirk Schadendorf, Philipp Jansen, Cindy Franklin, Tim Holland-Letz, Dieter Krahl,

et al. Pathologist-level classification of histopathological melanoma images with deep neural networks. *European Journal of Cancer*, 115:79–83, 2019.

[30] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[31] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[32] Humayun Irshad, Ludovic Roux, and Daniel Racoceanu. Multi-channels statistical and morphological features based mitosis detection in breast cancer histopathology. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6091–6094. IEEE, 2013.

[33] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.

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

[35] Stevan R Knezevich, Raymond L Barnhill, David E Elder, Michael W Piepkorn, Lisa M Reisch, Gaia Pocobelli, Patricia A Carney, and Joann G Elmore. Variability in mitotic figures in serial sections of thin melanomas. *Journal of the American Academy of Dermatology*, 71(6):1204–1211, 2014.

[36] Carol L Kosary, Sean F Altekruse, Jennifer Ruhl, Richard Lee, and Lois Dickie. Clinical and prognostic factors for melanoma of the skin using seer registries: collaborative stage data collection system, version 1 and version 2. *Cancer*, 120:3807–3814, 2014.

[37] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[38] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2021.

[39] Jing Li and Nigel M Allinson. A comprehensive review of current local features for computer vision. *Neurocomputing*, 71(10):1771–1787, 2008.

[40] Yuguang Li, Ezgi Mercan, Stevan Knezevitch, Joann G Elmore, and Linda G Shapiro. Efficient and accurate mitosis detection-a lightweight rcnn approach. In *ICPRAM*, pages 69–77, 2018.

[41] Zeju Li, Konstantinos Kamnitsas, and Ben Glocker. Overfitting of neural nets under class imbalance: Analysis and improvements for segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 402–410. Springer, 2019.

[42] Santiago López-Tapia, José Aneiros-Fernández, and Nicolás Pérez de la Blanca. A fast pyramidal bayesian model for mitosis detection in whole-slide images. In *European Congress on Digital Pathology*, pages 135–143. Springer, 2019.

[43] Cheng Lu and Mrinal Mandal. Automated segmentation and analysis of the epidermis area in skin histopathological images. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pages 5355–5359. IEEE, 2012.

[44] Cheng Lu and Mrinal Mandal. Automated analysis and diagnosis of skin melanoma on whole slide histopathological images. *Pattern Recognition*, 48(8):2738–2750, 2015.

[45] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

[46] Dwarikanath Mahapatra, Behzad Bozorgtabar, Jean-Philippe Thiran, and Mauricio Reyes. Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 580–588. Springer, 2018.

[47] Sébastien Marcel and Yann Rodriguez. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1485–1488, 2010.

[48] Anne L Martel, Dan Hosseinzadeh, Caglar Senaras, Yu Zhou, Azadeh Yazdanpanah, Rushin Shojaii, Emily S Patterson, Anant Madabhushi, and Metin N Gurcan. An image analysis resource for cancer research: Piip—pathology image informatics platform for visualization, analysis, and management. *Cancer research*, 77(21):e83–e86, 2017.

[49] Sachin Mehta, Ezgi Mercan, Jamen Bartlett, Donald Weaver, Joann Elmore, and Linda Shapiro. Learning to segment breast biopsy whole slide images. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 663–672. IEEE, 2018.

[50] Sachin Mehta, Mohammad Rastegari, Anat Caspi, Linda Shapiro, and Hannaneh Hajishirzi. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In *ECCV*, 2018.

[51] Caner Mercan, Selim Aksoy, Ezgi Mercan, Linda G Shapiro, Donald L Weaver, and Joann G Elmore. Multi-instance multi-label learning for multi-class classification of whole slide breast histopathology images. *IEEE transactions on medical imaging*, 37(1):316–325, 2017.

[52] Caner Mercan, Bulut Aygunes, Selim Aksoy, Ezgi Mercan, Linda G Shapiro, Donald L Weaver, and Joann G Elmore. Deep feature representations for variable-sized regions of interest in breast histopathology. *IEEE Journal of Biomedical and Health Informatics*, 25(6):2041–2049, 2020.

[53] Ezgi Mercan, Sachin Mehta, Jamen Bartlett, Linda G Shapiro, Donald L Weaver, and Joann G Elmore. Assessment of machine learning of breast pathology structures for automated differentiation of breast cancer and high-risk proliferative lesions. *JAMA network open*, 2(8):e198777–e198777, 2019.

[54] Minitab. *Minitab Blog*, 2015.

[55] Zahra Mirikharaji and Ghassan Hamarneh. Star shape prior in fully convolutional networks for skin lesion segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 737–745. Springer, 2018.

[56] NationalCancerInstitute. National cancer institute. `https://www.cancer.gov/about-cancer/understanding/what-is-cancer#types`, 2021. (Accessed on 05/10/2021).

[57] Haomiao Ni, Hong Liu, Kuansong Wang, Xiangdong Wang, Xunjian Zhou, and Yueliang Qian. Wsi-net: Branch-based and hierarchy-aware network for segmentation and classification of breast histopathological whole-slide images. In *International Workshop on Machine Learning in Medical Imaging*, pages 36–44. Springer, 2019.

[58] Shima Nofallah, Sachin Mehta, Ezgi Mercan, Stevan Knezevich, Caitlin J May, Donald Weaver, Daniela Witten, Joann G Elmore, and Linda Shapiro. Machine learning techniques for mitoses classification. *Computerized Medical Imaging and Graphics*, 87:101832, 2021.

[59] Shima Nofallah, Mojgan Mokhtari, Wenjun Wu, Sachin Mehta, Stevan Knezevich, Caitlin J May, Oliver H Chang, Annie C Lee, Joann G Elmore, and Linda G Shapiro.

Segmenting skin biopsy images with coarse and sparse annotations using u-net. *Journal of Digital Imaging*, pages 1–12, 2022.

[60] C Ortiz de Solorzano, R Malladi, SA Lelievre, and SJ Lockett. Segmentation of nuclei and cells using membrane related protein markers. *journal of Microscopy*, 201(3):404–415, 2001.

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

[62] Anabik Pal, Utpal Garain, Aditi Chandra, Raghunath Chatterjee, and Swapan Senapati. Psoriasis skin biopsy image segmentation using deep convolutional neural network. *Computer methods and programs in biomedicine*, 159:59–69, 2018.

[63] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[64] Sokol Petushi, Fernando U Garcia, Marian M Haber, Constantine Katsinis, and Aydin Tozeren. Large-scale computations on histology images reveal grade-differentiating parameters for breast cancer. *BMC medical imaging*, 6(1):14, 2006.

[65] Michael W Piepkorn, Raymond L Barnhill, David E Elder, Stevan R Knezevich, Patricia A Carney, Lisa M Reisch, and Joann G Elmore. The mpath-dx reporting schema for melanocytic proliferations and melanoma. *Journal of the American Academy of Dermatology*, 70(1):131–141, 2014.

[66] Ronaldo C Prati, Gustavo EAPA Batista, and Maria Carolina Monard. Data mining with imbalanced class distributions: concepts and methods. In *IICAI*, pages 359–376, 2009.

[67] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

[68] Dezső Ribli, Anna Horváth, Zsuzsa Unger, Péter Pollner, and István Csabai. Detecting and classifying lesions in mammograms with deep learning. *Scientific reports*, 8(1):1–7, 2018.

[69] Darrell S Rigel and John A Carucci. Malignant melanoma: prevention, early detection, and treatment in the 21st century. *CA: a cancer journal for clinicians*, 50(4):215–236, 2000.

[70] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[71] Vincent Roullier, Olivier Lézoray, Vinh-Thong Ta, and Abderrahim Elmoataz. Mitosis extraction in breast-cancer histopathological whole slide images. In *International Symposium on Visual Computing*, pages 539–548. Springer, 2010.

[72] Ludovic Roux, Daniel Racoceanu, Nicolas Loménie, Maria Kulikova, Humayun Irshad, Jacques Klossa, Frédérique Capron, Catherine Genestie, Gilles Le Naour, and Metin N Gurcan. Mitosis detection in breast cancer histological images an icpr 2012 contest. *Journal of pathology informatics*, 4, 2013.

[73] Monjoy Saha, Chandan Chakraborty, and Daniel Racoceanu. Efficient deep learning model for mitosis detection using breast histopathology images. *Computerized Medical Imaging and Graphics*, 64:29–40, 2018.

[74] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[75] Adnan Saood and Iyad Hatem. Covid-19 lung ct image segmentation using deep learning methods: U-net versus segnet. *BMC Medical Imaging*, 21(1):1–10, 2021.

[76] Olcay Sertel, Umit V Catalyurek, Hiroyuki Shimada, and Metin N Gurcan. Computer-aided prognosis of neuroblastoma: Detection of mitosis and karyorrhexis cells in digitized histological images. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1433–1436. IEEE, 2009.

[77] Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, Pascal Fua, and Francesc Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 118–126, 2015.

[78] Korsuk Sirinukunwattana, Shan E Ahmed Raza, Yee-Wah Tsang, David RJ Snead, Ian A Cree, and Nasir M Rajpoot. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE transactions on medical imaging*, 35(5):1196–1206, 2016.

[79] Jamshid Sourati, Ali Gholipour, Jennifer G Dy, Sila Kurugol, and Simon K Warfield. Active deep learning with fisher information for patch-wise semantic segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 83–91. Springer, 2018.

[80] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, 63:101693, 2020.

[81] John F Thompson, Seng-Jaw Soong, Charles M Balch, Jeffrey E Gershenwald, Shouluan Ding, Daniel G Coit, Keith T Flaherty, Phyllis A Gimotty, Timothy Johnson, Marcella M Johnson, et al. Prognostic significance of mitotic rate in localized primary cutaneous

melanoma: an analysis of patients in the multi-institutional american joint committee on cancer melanoma staging database. *Journal of Clinical Oncology*, 29(16):2199, 2011.

[82] Mike Van Zon, Nikolas Stathonikos, Willeke AM Blokx, Selim Komina, Sybren LN Maas, Josien PW Pluim, Paul J Van Diest, and Mitko Veta. Segmentation and classification of melanoma and nevus in whole slide images. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 263–266. IEEE, 2020.

[83] Mitko Veta, Josien PW Pluim, Nikolaos Stathonikos, Paul J van Diest, Francisco Beca, and Andrew Beck. Tumor proliferation assessment challenge 2016, miccai grand challenge. 2016.

[84] 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.

[85] Jelte Peter Vink and Gerard de Haan. No-reference metric design with machine learning for local video compression artifact level. *IEEE Journal of Selected Topics in Signal Processing*, 5(2):297–308, 2011.

[86] JP Vink, MB Van Leeuwen, CHM Van Deurzen, and G De Haan. Efficient nucleus detector in histopathology images. *Journal of microscopy*, 249(2):124–135, 2013.

[87] Carolina Wählby, I-M SINTORN, Fredrik Erlandsson, Gunilla Borgefors, and Ewert Bengtsson. Combining intensity, edge and shape information for 2d and 3d segmentation of cell nuclei in tissue sections. *Journal of microscopy*, 215(1):67–76, 2004.

[88] Haibo Wang, Angel Cruz-Roa, Ajay Basavanhally, Hannah Gilmore, Natalie Shih, Mike Feldman, John Tomaszewski, Fabio Gonzalez, and Anant Madabhushi. Cascaded ensemble of convolutional neural networks and handcrafted features for mitosis detection.

In *Medical Imaging 2014: Digital Pathology*, volume 9041, page 90410B. International Society for Optics and Photonics, 2014.

[89] Ying Wang, Ping Han, Xiaoguang Lu, Renbiao Wu, and Jingxiong Huang. The performance comparison of adaboost and svm applied to sar atr. In *Radar, 2006. CIE'06. International Conference on*, pages 1–4. IEEE, 2006.

[90] Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

[91] Wenjun Wu, Sachin Mehta, Shima Nofallah, Stevan Knezevich, Caitlin J May, Oliver H Chang, Joann G Elmore, and Linda G Shapiro. Scale-aware transformers for diagnosing melanocytic lesions. *IEEE Access*, 9:163526–163541, 2021.

[92] Ellery Wulczyn, David F Steiner, Zhaoyang Xu, Apaar Sadhwani, Hongwu Wang, Isabelle Flament-Auvigne, Craig H Mermel, Po-Hsuan Cameron Chen, Yun Liu, and Martin C Stumpe. Deep learning-based survival prediction for multiple cancer types using histopathology images. *PloS one*, 15(6):e0233678, 2020.

[93] Hongming Xu, Cheng Lu, Richard Berendt, Naresh Jha, and Mrinal Mandal. Automated analysis and classification of melanocytic tumor on skin whole slide images. *Computerized medical imaging and graphics*, 66:124–134, 2018.

[94] Hongming Xu and Mrinal Mandal. Epidermis segmentation in skin histopathological images based on thickness measurement and k-means algorithm. *EURASIP Journal on Image and Video Processing*, 2015(1):1–14, 2015.

[95] Pavel Yakubovskiy. Segmentation models pytorch. `https://github.com/qubvel/segmentation_models.pytorch`, 2020.

[96] Ling Zhang, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie Harmon, Baris Turkbey, Holger Roth, Andriy Myronenko, Daguang Xu, and Ziyue Xu. When unseen

domain generalization is unnecessary? rethinking data augmentation. *arXiv preprint arXiv:1906.03347*, 2019.

[97] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018.

[98] Yishuo Zhang and Albert CS Chung. Deep supervision with additional labels for retinal vessel segmentation task. In *International conference on medical image computing and computer-assisted intervention*, pages 83–91. Springer, 2018.

[99] Hao Zheng, Lin Yang, Jianxu Chen, Jun Han, Yizhe Zhang, Peixian Liang, Zhuo Zhao, Chaoli Wang, and Danny Z Chen. Biomedical image segmentation via representative annotation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5901–5908, 2019.

[100] Yao Zhou, Hua Mao, and Zhang Yi. Cell mitosis detection using deep neural networks. *Knowledge-Based Systems*, 137:19–28, 2017.

VITA

Shima Nofallah received her B.Sc. and M.Sc. degrees in Biomedical Engineering from Amirkabir University of Technology in Iran. She is currently a Ph.D. candidate in Electrical and Computer Engineering at the University of Washington. Her research interests include computer vision, machine learning, and medical image processing.