A Pathology Deep Learning System Capable of Triage of Melanoma Specimens Utilizing Dermatopathologist Consensus as Ground Truth *

Sivaramakrishnan Sankarapandian1, Saul Kohn1, Vaughn Spurrier1, Sean Grullon1, Rajath E. Soans1, Kameswari D. Ayyagari1, Ramachandra V. Chamarthi1, Kiran Motaparthi2, Jason B. Lee3, Wonwoo Shon4, Michael Bonham1, and Julianna D. Ianni1

1Proscia Inc.
2Department of Dermatology, University of Florida College of Medicine
3Department of Dermatology, Sidney Kimmel Medical College at Thomas Jefferson University
4Department of Pathology and Laboratory Medicine, Cedars-Sinai

Abstract

Although melanoma occurs more rarely than several other skin cancers, patients’ long term survival rate is extremely low if the diagnosis is missed. Diagnosis is complicated by a high discordance rate among pathologists when distinguishing between melanoma and benign melanocytic lesions. A tool that allows pathology labs to sort and prioritize melanoma cases in their workflow could improve turnaround time by prioritizing challenging cases and routing them directly to the appropriate subspecialist. We present a pathology deep learning system (PDLS) that performs hierarchical classification of digitized whole slide image (WSI) specimens into six classes defined by their morphological characteristics, including classification of “Melanocytic Suspect” specimens likely representing melanoma or severe dysplastic nevi. We trained the system on 7,685 images from a single lab (the reference lab), including the largest set of triple-concordant melanocytic specimens compiled to date, and tested the system on 5,099 images from two distinct validation labs. We achieved Area Underneath the ROC Curve (AUC) values of 0.93 classifying Melanocytic Suspect specimens on the reference lab, 0.95 on the first validation lab, and 0.82 on the second validation lab. We demonstrate that the PDLS is capable of automatically sorting and triaging skin specimens with high sensitivity to Melanocytic Suspect cases and that a pathologist would only need between 30% and 60% of the caseload to address all melanoma specimens.

1. Introduction

More than 5 million diagnoses of skin cancer are made each year in the United States, about 106,000 of which are melanoma of the skin [1]. Diagnosis requires microscopic examination of hematoxylin and eosin (H&E) stained, paraffin wax embedded biopsies of skin lesion specimens on glass slides. These slides can be manually observed under a microscope, or digitally on a whole slide image (WSI) scanned on specialty hardware.

The 5-year survival rate of patients with metastatic malignant melanoma is less than 20% [2]. Melanoma occurs more rarely than several other types of skin cancer, and its diagnosis is challenging, as evidenced by a high discordance rate among pathologists when distinguishing between melanoma and benign melanocytic lesions (∼ 40% discordance rate; e.g. [3, 4]). The Melanocytic Pathology Assessment Tool and Hierarchy for Diagnosis (MPATH-Dx; MPATH hereafter) reporting schema was introduced by Piepkorn et al. [5] to provide a precise and consistent framework for dermatopathologists to grade the severity of melanocytic proliferation in a specimen. MPATH scores are enumerated from I to V, with I denoting a benign melanocytic lesion and V denoting invasive melanoma. It has been shown that discordance rates are related to the MPATH score, with better inter-observer agreement on both ends of the scale than in the middle (e.g. [4]).

A tool that allows labs to sort and prioritize melanoma cases in advance of pathologist review could improve turnaround time, enabling pathologists to review cases re-
quiring faster turnaround time early in the day. This is particularly important as shorter turnaround time is correlated with improved overall survival for melanoma patients [9]. It could also alleviate common lab bottlenecks such as referring cases to specialized dermatopathologists, or ordering additional tissue staining beyond the standard H&E. These contributions are especially important as the number of skin biopsies performed per year has skyrocketed, while the number of practicing pathologists has declined [7].

The advent of digital pathology has brought the revolution in machine learning and artificial intelligence to bear on a variety of tasks common to pathology labs. Several deep learning algorithms have been introduced to distinguish between different skin cancers and healthy tissue with very high accuracy (e.g. [8–11]). However, almost all of these studies fail to demonstrate the robustness required for use in a clinical workflow setting setting because they were tested on a small number (<1000) of WSIs. Moreover, these algorithms are often not capable of triaging WSIs, as they use curated training and test datasets that do not represent the diversity of cases encountered in a dermatopathology lab. Many of them rely on pixel-level annotations to train their models, which is slow and expensive to scale to a large dataset with greater variability. Considerable advancements have been made towards systems capable of use in clinical practice for prostate cancer. Campanella et al. [12] trained a model in a weakly-supervised framework that did not require pixel-level annotations to classify prostate cancer and validated on ~10,000 WSIs sourced from multiple countries. However, some degree of human-in-the-loop curation was performed on their dataset, including manual quality control such as post-hoc removal of slides with pen ink from the study. Pantanowitz et al. [13] used pixel-wise annotations to develop a model trained on ~550 WSIs that distinguishes high-grade from low-grade prostate cancer. In dermatopathology, the model developed in [14] classified skin lesion specimens between four morphology-based groups, was tested on ~13,500 WSIs, and also demonstrated that use of confidence thresholding could provide a high accuracy; however, it grouped malignant melanoma with all other benign melanocytic lesions, limiting its potential uses. Additionally, all previous attempts at pathology classification using deep learning have, at their greatest level of abstraction, performed classification at the level of a WSI or a sub-region of a WSI. Since a pathologist is required to review all WSIs from a tissue specimen, previous deep learning pathology efforts therefore do not leverage the same visual information that a pathologist would have at hand to perform a diagnosis, require some curation of datasets to ensure that pathology is present in all training slides, and implement ad-hoc rules for combining the predictions of each WSI corresponding to a specimen. Most have also neglected the effect of diagnostic discordance on their ground truth, resulting in potentially mislabeled training and testing data.

In this work, we present a pathology deep learning system (PDLS) that can classify skin cases for triage and prioritization prior to pathologist review. Unlike previous systems, our PDLS performs hierarchical melanocytic specimen classification into low (MPATH I-II), Intermediate (MPATH III), or High (MPATH IV-V) diagnostic categories, enabling prioritization of melanoma cases. Our PDLS is the first to classify skin biopsies at the specimen level through a collection of WSIs that represent the entirety of the tissue from a single specimen. This training procedure ultimately models the process of a dermatopathologist, who reviews the full collection of scanned WSIs corresponding to a specimen to make a diagnosis. Finally, we train and validate the PDLS on the largest dataset of consensus-reviewed melanocytic specimens published to date. The system is built to be scalable and ready for the real-world, built without any pixel-level annotations, and incorporating the automatic removal of scanning artifacts.

2. Methods

2.1. Reference and Validation Lab Data Collection

The PDLS was trained using slides from 3511 specimens (consisting of 7685 WSIs) collected from a leading dermatopathology lab in a top academic medical center (Department of Dermatology at University of Florida College of Medicine), which we refer to as the ‘Reference Lab’. The Reference Lab dataset consisted of both an uninterrupted series of sequentially-accessioned cases (69% of total specimens) and a targeted set, curated to enrich for rarer melanocytic pathologies (31% of total specimens). Melanocytic specimens were only included in this set if 3 dermatopathologists’ consensus on diagnosis could be established. The WSIs consisted exclusively of H&E-stained, formalin-fixed, paraffin-embedded dermatopathology tissue and were scanned using a 3DHistech P250 High Capacity Slide Scanner at an objective power of 20X, corresponding to 0.24µm/pixel. The final classification given by the PDLS was one of six classes, defined by their morphologic characteristics —

1. Basaloid: containing abnormal proliferations of basaloid-oval cells, primarily basal cell carcinoma of various types
2. Squamous: containing malignant squamoid epithelial proliferations, consisting primarily of squamous cell carcinoma (invasive and in situ)
3. Melanocytic Low Risk: benign to moderately atypical melanocytic nevi/proliferation of cells of melanocytic origin, classified as the MPATH I or MPATH II diagnostic category...
Table 1: Counts of each of the general pathologies in the reference set, broken-out into specific diagnostic entities.

<table>
<thead>
<tr>
<th>Diagnostic Morphology</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Basaloid</strong></td>
<td>544</td>
</tr>
<tr>
<td>Nodular Basal Cell Carcinoma</td>
<td>404</td>
</tr>
<tr>
<td>Basal Cell Carcinoma, NOS</td>
<td>123</td>
</tr>
<tr>
<td>Basal Cell Carcinoma, Morphea type</td>
<td>7</td>
</tr>
<tr>
<td>Pilomatrixoma</td>
<td>5</td>
</tr>
<tr>
<td>Infiltrative Basal Cell Carcinoma</td>
<td>5</td>
</tr>
<tr>
<td><strong>Squamous</strong></td>
<td>530</td>
</tr>
<tr>
<td>Invasive Squamous Cell Carcinoma</td>
<td>269</td>
</tr>
<tr>
<td>Squamous Cell Carcinoma in situ</td>
<td>254</td>
</tr>
<tr>
<td>(Bowen’s Disease)</td>
<td></td>
</tr>
<tr>
<td>Fibrokeratoma</td>
<td>4</td>
</tr>
<tr>
<td>Warty Dyskeratoma</td>
<td>3</td>
</tr>
<tr>
<td><strong>Melanocytic High Risk</strong></td>
<td>102</td>
</tr>
<tr>
<td>Melanoma</td>
<td>102</td>
</tr>
<tr>
<td><strong>Melanocytic Intermediate Risk</strong></td>
<td>213</td>
</tr>
<tr>
<td>Melanoma In Situ</td>
<td>202</td>
</tr>
<tr>
<td>Severe Dysplasia</td>
<td>9</td>
</tr>
<tr>
<td><strong>Melanocytic Low Risk</strong></td>
<td>764</td>
</tr>
<tr>
<td>Conventional Melanocytic Nevus</td>
<td>368</td>
</tr>
<tr>
<td>(acquired and congenital)</td>
<td></td>
</tr>
<tr>
<td>Mild Dysplasia</td>
<td>289</td>
</tr>
<tr>
<td>Moderate Dysplasia</td>
<td>75</td>
</tr>
<tr>
<td>Halo Nevus</td>
<td>14</td>
</tr>
<tr>
<td>Dysplastic Nevus, NOS</td>
<td>12</td>
</tr>
<tr>
<td>Spitz Nevus</td>
<td>2</td>
</tr>
<tr>
<td>Blue Nevus</td>
<td>2</td>
</tr>
<tr>
<td><strong>Other Diagnoses</strong></td>
<td>1360</td>
</tr>
</tbody>
</table>

4. Melanocytic Intermediate Risk: severely atypical melanocytic nevi or melanoma in situ, classified as the MPATH III diagnostic category.

5. Melanocytic High Risk: invasive melanoma, classified as the MPATH IV or V diagnostic category.

6. Other: all skin specimens that do not fit into the above classes, including but not limited to inflammatory conditions and benign proliferations of squamoid epithelial cells.

The overall reference set was composed of 544 Basaloid, 530 Squamous, 1079 Melanocytic and 1358 Other specimens. Of the Melanocytic specimens, 764 were Low Risk, 213 were Intermediate Risk and 102 were High Risk. The heterogeneity of this reference set is illustrated in Table 1.

The specimen counts that we report for the melanocytic classes throughout this work reflect counts following three-way consensus review (see Section 2.2). For training, validating, and testing the PDLS, we divided this dataset into three partitions by sampling at random without replacement with 70% of specimens used for training, and 15% used for each of validation and testing.

Table 2: Class counts for the Validation Lab datasets. (The counts for the Reference Lab are given in Table 1.)

<table>
<thead>
<tr>
<th>Label category</th>
<th>Validation Lab 1</th>
<th>Validation Lab 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPATH I-II</td>
<td>1457</td>
<td>458</td>
</tr>
<tr>
<td>MPATH III</td>
<td>225</td>
<td>364</td>
</tr>
<tr>
<td>MPATH IV-V</td>
<td>100</td>
<td>361</td>
</tr>
<tr>
<td>Basaloid</td>
<td>198</td>
<td>265</td>
</tr>
<tr>
<td>Squamous</td>
<td>104</td>
<td>55</td>
</tr>
<tr>
<td>Other</td>
<td>711</td>
<td>563</td>
</tr>
</tbody>
</table>

Table 2 shows the class distribution for the Validation labs.

2.2. Consensus Review

There are high discordance rates in diagnosing melanocytic specimens. Elmore et al. [4] studied 240 dermatopathology cases and found that the consensus rate for MPATH Class II lesions was 25%, for MPATH Class III lesions 40%, and for MPATH Class IV 45%. Therefore, three board-certified pathologists reviewed each melanocytic specimen to establish a reliable ground truth for melanocytic cases in our study. The first review was the original specimen diagnosis made via glass slide examination under a microscope. Two additional dermatopathologists independently reviewed and rendered a diagnosis dig-
review. Melanocytic specimens were considered to have a birth and gender were provided with each specimen upon itally for each melanocytic specimen. The patient's year of we adopted the image adaptation procedure in [14].

The colors of the training data and validation data [18–20], each tile [16, 17]. In order to avoid domain shift between by setting a threshold on the variance of the Laplacian over inference on the test set. We also sought to remove areas removed from the training and validation data and before revisited model trained to detect inked slides. These tiles were identified by a weakly super-

highly correlated with malignant or High Risk pathologies.
sented a biased distractor signal in training the PDLS that is presented with three tasks: differentiating “Melanocytic Suspect” from ”Rest” class. Specimens that were classified as “Melanocytic Suspect” were fed into a downstream model, which further classified the specimen between “Melanocytic High Risk, Melanocytic Intermediate Risk” and “Rest”. The remaining specimens, classified as “Rest”, were fed into a separate downstream model, which further classified the specimen between “Basaloid, Squamous, Melanocytic Low Risk” and “Other”. Each model consisted of four fully-connected layers (two layers of 1024 channels each, followed by two of 512 channels each). Each neuron in the three layers after the input layer was ReLU ac-
tivated. The three models in the hierarchy were trained under a weakly-supervised multiple-instance learning (MIL) paradigm. Each embedded tile was treated as an instance of a bag containing all quality-assured tiles of a specimen. Embedded tiles were aggregated using sigmoid-activated attention heads [23][24]. To help prevent over-fitting, the training dataset consisted of augmented versions (inspired by Tellez et al. [25]) of the tiles. Augmentations were generated with the following augmentation strategies: random variations in brightness, hue, contrast, saturation, (up to a maximum of 15%), Gaussian noise with 0.001 variance, and random 90 degree image rotations. The upstream binary “Melanocytic Suspect vs. Rest” classification model and the downstream “Rest” subclassifier were each trained end-to-end with cross-entropy loss. The “Melanocytic Suspect” subclassifier was also trained with cross-entropy loss, but with a multi-task learning strategy. This subclassifier was presented with three tasks: differentiating “Melanocytic High Risk” from “Melanocytic Intermediate Risk” specimens, “Melanocytic High Risk” from “Rest” specimens, and “Melanocytic Intermediate Risk” from “Rest” specimens. The training loss for this subclassifier was computed for each task, but was masked if it did not relate to the ground truth label of the specimen. Two out of three tasks were trained for any given specimen in a training batch. By training in this manner, we were able to use the shared network layers as a generic representation of melanocytic pathologies, while the task branches learned to attend to specific differences to accomplish their tasks.

A diagnosis was rendered in the above fashion for every melanocytic specimen obtained from the Reference Lab and Validation Lab 1. All dysplastic and malignant melanocytic specimens from Validation Lab 2 were reviewed by three dermatopathologists and only the specimens for which consensus could be established were included in the study. No non-melanocytic specimens were reviewed for concordance due to inherently lower known rates of discordance [15].

For the specimens obtained from the Reference Lab, consensus was established for 75% of specimens originally diagnosed as MPATH I/II, 66% of those diagnosed as MPATH III, 87% of those diagnosed as MPATH IV/V, and for 74% of the reviewed specimens in total. For specimens obtained from Validation Lab 1, pathologists consensus was established for 84% of specimens originally diagnosed as MPATH I/II specimens, 51% of those diagnosed as MPATH III, 54% of those diagnosed as MPATH IV/V, and for 61% of the reviewed specimens in total.

2.3. Pathology Deep Learning System Architecture

The PDLS consists of three main components: quality control, feature extraction and hierarchical classification. A diagram of the PDLS is shown in Figure[1]. Each specimen was first segmented into tissue-containing regions, subdivided into 128x128 pixel tiles, and extracted at an objective power of 10X. Each tile was passed through the quality control and embedding components of the PDLS.

Quality control consisted of ink filtering, blur filtering, and image adaptation. Pen ink is common in labs migrating their workload from glass slides to WSIs where the location of possible malignancy was marked. This pen ink represented a biased distractor signal in training the PDLS that is highly correlated with malignant or High Risk pathologies. Tiles containing pen ink were identified by a weakly supervised model trained to detect inked slides. These tiles were removed from the training and validation data and before inference on the test set. We also sought to remove areas of the image that were out of focus due to scanning errors by setting a threshold on the variance of the Laplacian over each tile [16][17]. In order to avoid domain shift between the colors of the training data and validation data [18–20], we adopted the image adaptation procedure in [14].

The next component of the PDLS extracted informative features from the quality controlled, color-standardized tiles. To capture higher-level features in these tiles, they were propagated through a neural network (ResNet50: [21]) trained on the ImageNet [22] dataset to embed each input tile into 1024 channel vectors which were then used in our subsequent models.

We developed a hierarchical model architecture in order to classify both Melanocytic High and Intermediate Risk specimens with high sensitivity. First, an upstream model performed a binary classification between “Melanocytic Suspect” (defined as “High or Intermediate Risk”) and “Basaloid, Squamous, Low Risk” or “Other” (which we collectively define as the “Rest” class). Specimens that were classified as “Melanocytic Suspect” were fed into a downstream model, which further classified the specimen between “Melanocytic High Risk, Melanocytic Intermediate Risk” and “Rest”. The remaining specimens, classified as “Rest”, were fed into a separate downstream model, which further classified the specimen between “Basaloid, Squamous, Melanocytic Low Risk” and “Other”. Each model consisted of four fully-connected layers (two layers of 1024 channels each, followed by two of 512 channels each). Each neuron in the three layers after the input layer was ReLU ac-
tivated. The three models in the hierarchy were trained under a weakly-supervised multiple-instance learning (MIL) paradigm. Each embedded tile was treated as an instance of a bag containing all quality-assured tiles of a specimen. Embedded tiles were aggregated using sigmoid-activated attention heads [23][24]. To help prevent over-fitting, the training dataset consisted of augmented versions (inspired by Tellez et al. [25]) of the tiles. Augmentations were generated with the following augmentation strategies: random variations in brightness, hue, contrast, saturation, (up to a maximum of 15%), Gaussian noise with 0.001 variance, and random 90 degree image rotations. The upstream binary “Melanocytic Suspect vs. Rest” classification model and the downstream “Rest” subclassifier were each trained end-to-end with cross-entropy loss. The “Melanocytic Suspect” subclassifier was also trained with cross-entropy loss, but with a multi-task learning strategy. This subclassifier was presented with three tasks: differentiating “Melanocytic High Risk” from “Melanocytic Intermediate Risk” specimens, “Melanocytic High Risk” from “Rest” specimens, and “Melanocytic Intermediate Risk” from “Rest” specimens. The training loss for this subclassifier was computed for each task, but was masked if it did not relate to the ground truth label of the specimen. Two out of three tasks were trained for any given specimen in a training batch. By training in this manner, we were able to use the shared network layers as a generic representation of melanocytic pathologies, while the task branches learned to attend to specific differences to accomplish their tasks.
Figure 1: The stages of the PDLS are: Quality Control, Feature Extraction, and Hierarchical Classification. All single specimen WSIs were first passed through the tiling stage, then the quality control stage consisting of ink filtering, blur filtering and image adaptation. The image-adapted tiles were passed through the feature extraction stage consisting of a pre-trained ResNet50 network to obtain embedded vectors. Finally, the vectors were propagated through the hierarchical classification stage consisting of an upstream model performing a binary classification between “Melanocytic Suspect” and “Rest”. Specimens that were classified as “Melanocytic Suspect” were fed into a downstream model, which classified between “Melanocytic High Risk, Melanocytic Intermediate Risk” and “Rest”. The remaining specimens were fed into a separate downstream “Rest” model, which classified between “Basaloid, Squamous, Melanocytic Low Risk” and “Other”.

During inference, an input specimen’s predicted classes were computed as follows (see Figure 2):

1. The larger of the two confidence values (see below for our confidence thresholding procedure) output from the upstream classifier determined which downstream classifier a specimen was handed to.

2. If the specimen was handed to the “Rest” subclassifier, we used the highest confidence class probability as the predicted label.

3. If the specimen was handed to the Melanocytic Suspect subclassifier, we used the highest confidence class probability between the “Melanocytic High Risk vs. Rest” and “Melanocytic Intermediate Risk vs Rest” tasks as the predicted label.

As a final step in the classification pipeline, we performed classification with uncertainty quantification to establish a confidence score for each prediction using a Monte Carlo dropout method following the same procedure as used by Gal & Ghahramani [26]. Using the confidence distribution of the specimens in the validation set of the Reference Lab, we computed confidence threshold values for each predicted class following the procedure outlined Ianni et al. [14] by requiring classifications to meet a predefined level of accuracy in the validation set. Specimens that were predicted as “Melanocytic High Risk” had to pass two confidence thresholds: an accuracy threshold and a PPV threshold in order to be predicted as “Melanocytic High Risk”. Specimens that were predicted to be “Melanocytic High Risk” but failed to meet these thresholds were predicted as “Melanocytic Suspect”. We set thresholds that maximized the sensitivity of the PDLS to the “Melanocytic Suspect” class, while simultaneously maximizing the PPV to the “Melanocytic High Risk” class.

To evaluate how our PDLS generalizes to data from other labs, we fine-tuned the model trained on data from the Reference Lab to both Validation Lab 1 and Validation Lab 2. We specifically set aside 255 specimens from each validation lab (using an equal class distribution of specimens) as
Figure 2: The hierarchical model prediction procedure for a given input specimen. Specimens that were predicted to be “Melanocytic High Risk” had to pass both an accuracy threshold and a Positive Predictive Value (PPV) threshold (“PPV_t”) — both established a priori on the validation set — to be predicted as “Melanocytic High Risk”; otherwise they were predicted as “Melanocytic Suspect”. The confidence thresholding procedure is described in the text.

3. Results

3.1. PDSL Performance Evaluation

To demonstrate the performance of our PDSL, we first show Receiver Operating Characteristic (ROC) curves for the individual hierarchical component models. The ROC curves derived from the Reference Lab test dataset are shown in Figure 3. The Area Underneath the ROC Curve (AUC) values, calculated with the one-vs-rest scoring scheme, were 0.97, 0.95, 0.87, 0.84, 0.81, 0.93, and 0.96 for the Basaloid, Squamous, Other, Melanocytic High Risk, Melanocytic Intermediate Risk, Melanocytic Suspect, and Melanocytic Low Risk classes, respectively. Table 3 shows the PDSL performance with respect to diagnostic entities of clinical interest on the Reference lab test dataset. The sensitivity of the PDSL to the Melanocytic Suspect class was found to be 0.83, 0.85 for the Melanocytic High and Intermediate risk classes, respectively. The PPV to Melanocytic High Risk was found to be 0.57. The dropout Monte Carlo procedure set the threshold for Melanocytic High Risk classification very high; specimens below this threshold were classified as Melanocytic Suspect, maximizing the sensitivity to this class.

After fine-tuning all three models in the hierarchy through the calibration procedure in each validation lab, we report the ability of our PDSL to generalize to unseen data from both validation labs. Note that fine-tuning was not performed for any of the models in the pre-processing pipeline (Colorization, Ink Detection or ResNet). The ROC curves derived from the Validation Lab 1 and Validation Lab 2 test datasets are shown in Figure 3. The AUC values for Validation Lab 1 were 0.95, 0.88, 0.81, 0.87, 0.87, 0.95, and 0.92 for the Basaloid, Squamous, Other, Melanocytic High Risk, Intermediate Risk, Suspect, and Low Risk classes, respectively and the AUC values for the same classes for Validation Lab 2 were 0.93, 0.92, 0.69, 0.76, 0.75, 0.82, and 0.92.

3.2. Consensus Ablation Study

Diagnosing melanocytic cases is challenging. Although some specimens (such as ones diagnosed as Compound Nevi) clearly exhibit very low risk and others (such as invasive melanoma) exhibit very high risk of progressing into life threatening conditions, reproducible stratification in the middle of the morphological spectrum has historically proved difficult [4, 27]. The results were derived with our PDSL trained and evaluated on consensus data: data for which the ground truth melanocytic specimen diagnostic categories were agreed upon by multiple experts. To understand the effect of consensus on training deep learning models, we performed an ablation study by training two hierarchical models. Both models used all non-melanocytic specimens available in the training set. The first model was trained only including melanocytic specimens for which consensus was obtained under the diagnostic categories of MPATH I/II, MPATH III, or MPATH IV/V. The other model was trained by also including non-consensus data: melanocytic specimens whose diagnostic category was not agreed upon by the experts. To facilitate a fair comparison, we reserved separate validation sets for both model versions and a common consensus test set derived from the Reference Lab. The sensitivities of our PDSL to different classes on both consensus and non-consensus data are shown in Figure 4, where a a clear improvement is shown in the sensitivity to the Melanocytic class of over 40% for melanocytic specimens that are annotated with consensus labels over ones that are not; this primarily manifested from a reduction in false positive Melanocytic Suspect classifications.

4. Discussion

In this work, we demonstrated the performance of a PDSL capable of automatically sorting and triaging skin specimens with high sensitivity to Melanocytic Suspect cases prior to review by a pathologist. Existing methods (such as [23]) provide diagnostically-relevant information
Figure 3: Receiver Operating Characteristic (ROC) curves for the individual models - the upstream classifier (left column), the High & Melanocytic Intermediate classifier (middle column) and the Basaloid, Squamous, Low Risk Melanocytic & Rest classifier (right column) - from the hierarchical architecture of the PDLS are shown for the Reference Lab (first row), the first Validation Lab, (second row) and the second Validation Lab (third row).

on a potential melanoma specimen only after a pathologist has reviewed the specimen and classified it as a Melanocytic Suspect lesion. The PDLS’s ability to classify suspected melanoma prior to pathologist review could substantially reduce diagnostic turnaround time for melanoma by not only allowing timely review and expediting the ordering of additional tests or stains, but also ensuring that suspected melanoma cases are routed directly to subspecialists. The potential clinical impact of a PDLS with these capabilities is underscored by the fact that early melanoma detection is correlated with improved patient outcomes [6].

As the PDLS was optimized to maximize melanoma sensitivity, we investigated the performance of our PDLS as a simple Melanocytic Suspect binary classifier. We envisioned a scenario in which a pathologist’s work list of specimens is sorted by the system’s confidence (in descending order) in the upstream classifier’s suspect melanocytic classification. Figure 5 demonstrates the resulting sensitivity to the Melanocytic Suspect class against the percentage of total specimens that a pathologist would have to review in this scheme in order to achieve that sensitivity. In this dataset, we show that a pathologist would only need between 30% and 60% of the caseload to address all melanoma specimens.

Diagnostic classification of melanocytic lesions remains challenging. There is known lack of consensus among pathologists, and a disturbing lack of intra-pathologist concordance over time [29] was recently reported. Training with consensus data resulted in improved performance seen in classifications excluding Melanocytic Suspect, which has the highest pathologist discordance rates. Since pathologists tend to cautiously diagnose a benign lesion as malignant, our PDLS learned the same bias in absence of consensus. By training on consensus of multiple dermatopathologists, our PDLS may have the unique ability to learn a more consistent feature representation of melanoma and aid in flagging misdiagnosis. While the system is highly sensitive to melanoma (84% correctly detected as Intermedi-
Table 3: Metrics for selected diagnoses of clinical interest, based on the Reference Lab test set, representing the classification performance of the individual diagnoses into their higher-level classes: e.g., a correct classification of "Melanoma" is the prediction "Melanocytic High Risk". Results are class-weighted according to the relative prevalence in the test set.

Figure 4: Reference lab performance on the same test set when trained on consensus and non-consensus data. The melanocytic class is defined as the Low, Intermediate and High Risk classes. The sensitivity of the Melanocytic Intermediate and High Risk classes are defined with respect to the PDLS classifying these classes as suspect. The PPV to melanocytic high risk in the non-consensus trained model was 0.33, while the consensus model was 0.57.

Figure 5: Mean and standard deviation sensitivity to melanoma vs % reviewed for 1,000 simulated sequentially-accessioned datasets, drawn from Reference lab confidence scores. In the clinic, 95% of melanoma suspect cases are detected within the first 30% of cases, when ordered by melanoma suspect model confidence.

The PDLS system also enables other automated pathology workflows in addition to triage and prioritization of suspected melanoma cases. Sorting and triaging specimens into other classifications such as Basaloid could allow the majority of less complicated cases (such as basal cell carcinoma) to be directly assigned to general pathologists, or to dermatologists who routinely sign out such cases.

Relevant to any PDLS designed for clinical use is how well its performance generalizes to sites on which the system was not trained. Performance on the Validation Labs after calibration (as shown in Figure 3) was in many cases close to that of the Reference Lab. Further improvement is needed, however, in order to avoid the need for fine-tuning and improve model generalizability of previously trained models in new laboratory settings.
References


