Detection and Localization of Melanoma Skin Cancer in Histopathological Whole Slide Images

Neel Kanwal*[1], Roger Amundsen*[1], Helga Hardardottir†‡, Luca Tomasetti*, Erling Sandøy Undersrud†‡
Emiel A.M. Janssen†‡, Kjersti Engan*
*Department of Electrical Engineering and Computer Science, University of Stavanger, Stavanger, Norway
†Department of Chemistry, Bioscience and Environmental Engineering, University of Stavanger, Stavanger, Norway
‡Department of Pathology, Stavanger University Hospital, Stavanger, Norway

***Abstract*—If melanoma is diagnosed and treated in its early stages can increase the survival rate. A projected increase in skin cancer incidents and a shortage of dermatopathologists have emphasized the need for computational pathology (CPATH) systems. CPATH systems with deep learning (DL) models have the potential to identify the presence of melanoma by exploiting underlying morphological and cellular features. This paper proposes a DL method to detect melanoma and distinguish between normal skin and benign/malignant melanocytic lesions in whole slide images (WSI). Our method detects lesions with high accuracy and localizes them on a WSI to identify potential regions of interest for pathologists. The proposed method relies on using a single convolutional neural network to create localization maps first and use them to perform slide-level predictions to determine patients who have melanoma. Our best model provides favorable patch-wise classification results with a 0.992 F1 score and 0.99 sensitivity on unseen data. The source code is publicly available at Github.**

***Index Terms*—Computational Pathology, Deep Learning, Melanoma Diagnosis, Skin Cancer, Whole Slide Images**

I. Introduction

Malignant melanoma is an aggressive type of skin cancer [1]. Though melanoma only accounts for roughly 1% of skin cancer cases, it is the leading cause of mortality [2]. Melanoma skin cancer develops when melanocytes start to proliferate quickly and uncontrolled in the epidermis and dermis layers of the skin and form malignant lesions [2]. If not treated early, the tumor is likely to progress to another stage and can spread to other parts of the body [1]. Therefore, detection and accurate diagnosis at an early stage are of the utmost importance. Histopathological examination is a common practice for diagnosing cancer, where a skin sample is extracted by punch biopsy and processed to prepare a glass slide. Pathologists later use the glass slide to conduct microscopic inspection [3]. It is often time-consuming and challenging for humans to distinguish between early-stage cancer and benign lesions. A digitized version of a glass slide, known as a whole slide image, can overcome this hurdle of traditional histopathology by involving computational analysis [4]–[6]. Computational pathology (CPATH) systems using deep learning (DL) techniques can automate various diagnostic tasks by learning morphological and cellular patterns and providing predictions [6]–[9]. CPATH systems have a high potential to identify patients with melanoma and aid pathologists by providing a second opinion and localizing the regions of interest (ROIs).

Some previous works [11], [12] on melanoma diagnosis focused on selecting hand-crafted features from data such as first-order texture features, color intensity, entropy, eccentricity, etc. Jafari *et al.* [12] used dermoscopic images and obtained color-space transformation to extract morphological features. Their method used the ABCD rule (Asymmetry, Border irregularity, Color patterns, and Diameter) to distinguish between malignant and benign lesions. In an analogous approach, Sheha *et al.* [11] used gray-level co-occurrence matrix (GLCM) and texture features to discriminate between malignant melanoma and melanocytic nevi using a multilayer perceptron classifier. However, manual feature engineering is cumbersome, and data-driven features have been demonstrated to surpass classification performance in medical images [13], [14]. Although DL approaches automatically discover features from data, they require a significant amount of data and clinical labels, which might not always be available.

Transfer learning has been a common initialization strategy in DL to compensate for the lack of training data [3], [15]. It mainly relies on transferring knowledge to a DL model, trained on a different task, for performing a new classification task by automatically learning hidden features from the new data. Recently, Rajitha *et al.* [15] applied transfer learning over four convolutional neural networks (CNNs) to extract features and classify dermatophytosis in unstained skin samples. Their transfer learning method focused on freezing various fractions of feature extractors on a trial-and-error basis to find the best-performing architecture. In a similar CNN-based approach, Logu *et al.* [16] used histological images to differentiate between patches of cutaneous melanoma and healthy tissue. Their model was trained for a binary task and did not discriminate *benign lesions* from malignant lesions. Zhang *et al.* [17] extracted features at multiple scales and fused them to obtain a feature representation. They performed binary patch-wise classification but did not predict melanoma at the slide level. Wang *et al.* [18] detected malignant and non-

Corresponding author: neel.kanwal@uis.no.
[1] These authors contributed equally.

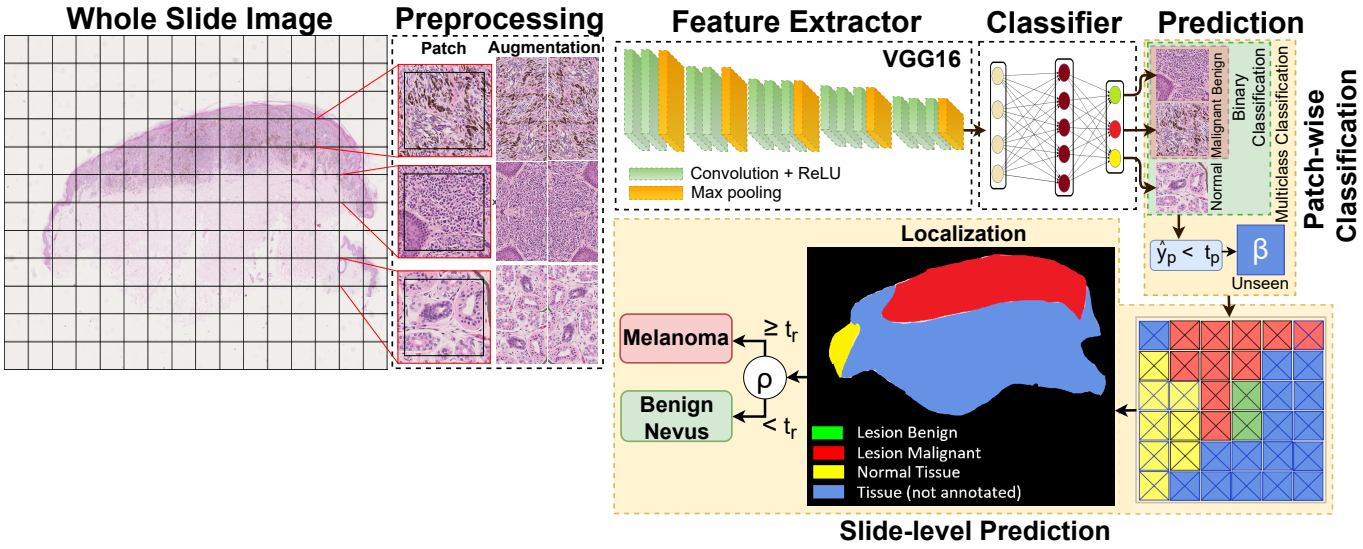

Figure 1. **An overview of our proposed melanoma detection method:** Whole slide images (WSIs) are divided into patches (sub-images) based on annotated regions. Preprocessing is performed to augment and normalize patches before feeding them to the feature extractor. VGG16 [10] is the backbone of our model, followed by the three-layer fully-connected classifier. Binary and multiclass models perform patch-wise classification and assign a class to each patch. Probability thresholding is applied to find patches of unseen tissue and produce output with one more (unseen) class. The classified patches are then combined using their coordinates to localize lesions. Finally, the localization maps are used for slide-level predictions of *melanoma* vs. *benign nevus* patients.

malignant nevi by CNNs and compared various architectures. Their experiments found VGG16 [10] superior in classification performance among other CNN architectures; However, their method relied on using a separate random forest classifier to perform slide-level predictions.

This paper proposes a CNN-based method to detect both *malignant* and *benign* lesions from *normal* and *unseen* tissue, as shown in Fig. 1, i.e., it provides a three-class prediction *(four-class output)* for each patch. The classified patches are then combined into segmentation maps to localize lesions in the WSI and reveal potential ROIs for a pathologist. Finally, the segmentation maps are used, with one extra learned parameter, to perform accurate slide-level classification as benign or malignant. Our method uses a single CNN network to perform both classification and segmentation tasks in one run.

II. Data Materials

We have analyzed 90 Hematoxylin and Eosin (H&E) stained skin biopsy whole slide images (WSIs) from Stavanger University Hospital (SUH) in Norway. These samples were clear benign or malignant, and no atypical nevi or in situ melanomas were included. The glass slides were scanned at 40x with Hamamatsu Nanozoomer s60 in *ndpi* format with a pixel size of 0.2199 µm x 0.2199 µm. All WSIs were provided with slide-level labels of *benign nevus*, *benign lentigo*, and *malignant melanoma* for 40, 7, and 43 WSIs, respectively. A split of 73/8/9 at the WSI level was carried out for training, validation, and testing. *Benign nevus* and *benign lentigo* were treated as a single label. A pathologist roughly annotated regions in each slide for benign lesions, malignant lesions, and other tissue types. By roughly, we mean that the annotator

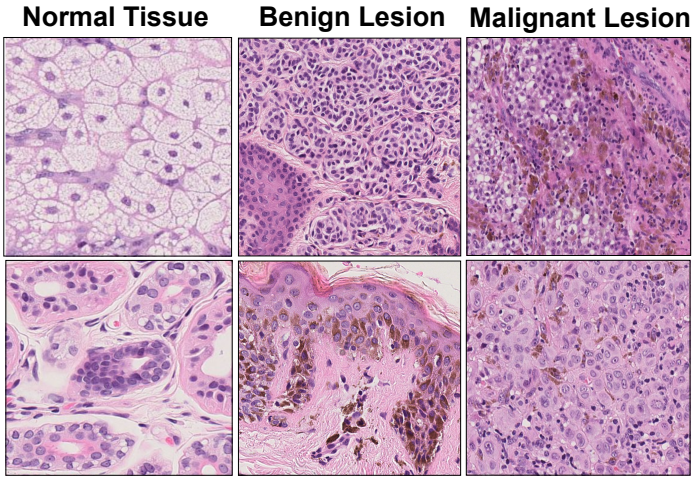

Figure 2. Sample patches of normal tissue, benign, and malignant lesion class from the dataset, extracted at 10x magnification.

should limit the time spent on each WSI. Annotated regions in WSIs are divided into small sub-images (patches) for a tractable computation [3], [9], [13]. Fig. 2 presents a few sample patches from each class extracted at 10x magnification level.

III. Proposed Method

A graphical overview of the melanoma detection method is presented in Fig. 1. Pretrained VGG16 [10] has been a popular choice for feature extractor for transfer learning in the literature [15], [18], [19], and our proposed method uses it as a

backbone. A custom classifier with three fully connected (FC) layers replaces the FC layers of VGG16. We proceed with the melanoma detection task by following a two-step approach. In the first step, we develop models for patch-wise classification in binary and multiclass fashion. The binary classification task distinguishes between malignant and benign lesion classes. The multiclass classification task uses three classes involving the normal tissue as well. In the second step, patches classified with a higher probability in a class are used to create a localization map and determine the slide-level outcome, as detailed below.

A. Pre-processing

Since the CNN can not handle the entire WSI at once due to the memory, the annotated regions in WSIs were split into small patches of 256×256 pixels at 2.5x, 10x, and 40x magnification levels. First, the background-foreground segmentation was performed by transforming RGB to HSV and thresholding Hue channel $\in [100 - 180]$ for purple and pink tones. Later, patches with a 70% overlap between foreground and annotation masks were extracted. We used the "patch-on-fly" for the memory-efficient patching process [9], [20]. To even-out class imbalance, geometric transformations with random crops of 224×224 were applied to the underrepresented class. Finally, all patches were resized to 224×224 pixels to match the input size of the pre-trained model before being normalized to the mean and standard deviation of the training set.

B. Patch-wise Classification

In the patch-wise classification step, we develop models in a binary and multiclass fashion that assign a single class to a patch. We trained a binary baseline model (Model$_{baseline}$) using a frozen feature extractor. Model$_{baseline}$ was trained on the dataset extracted at 10x magnification. Three binary models (Model$_{2.5x}$, Model$_{10x}$, and Model$_{40x}$) are trained with unfrozen feature extractors using dataset extracted at 2.5x, 10x, and 40x, respectively. Our multiclass model (Model$_{multiclass}$), with three output nodes, is trained on a 10x magnification dataset. For an input patch ($\boldsymbol{x}$) with the ground label (y_x), models output probability vector ($\boldsymbol{y}_p$) as shown in the Eq. (1), where $\hat{y}_B$, $\hat{y}_M$, and $\hat{y}_N$ are predicted probabilities for benign lesion, malignant lesion, and normal tissue class respectively.

$$\boldsymbol{y}_p = \begin{cases} [\hat{y}_B, \hat{y}_M] & \text{if Binary} \\ [\hat{y}_B, \hat{y}_M, \hat{y}_N] & \text{if Multiclass} \end{cases} \tag{1}$$

Since we use the entire WSI for the slide-level prediction step, the inference stage will encounter previously unseen tissue types. To handle the unseen tissue types, a new class (β) and probability threshold (t_p) are introduced. If the predicted output ($\boldsymbol{y}_p$) is less than t_p, then the predicted class ($\hat{y}_p$) is considered irrelevant tissue and assigned the β label as shown in Eq. (2). In short, the model output ($\boldsymbol{y}_p$) is either binary or with three classes, but the overall prediction ($\hat{y}_p$) includes one more class. Here, t_p helps to efficiently determine any lesion's presence without training models with new tissue classes. We determine the best t_p by maximizing lesion detection and minimizing false positives on the validation set.

$$\hat{y}_p = \begin{cases} argmax(\boldsymbol{y}_p) & \text{if } \; max(\boldsymbol{y}_p) \geq t_p \\ \beta, & \text{Otherwise} \end{cases} \tag{2}$$

C. Slide-level Prediction:

In the slide-level prediction step, we utilize the best-performing binary and multiclass models to create masks by putting together all classified patches from the previous step. Coordinates corresponding to the patch's location are used to fill the pixel in the down-sampled map with a color. Later, we calculate the ratio (ρ) of the malignant lesion class from the localization map using Eq. (3). We determine slide-level prediction ($\hat{y}_s$) of melanoma if ρ is greater than a ratio (t_r) (see Eq. (4)). In other words, t_r determines the degree of malignancy to report a patient with melanoma. A suitable value of t_r is selected based on the validation set to reduce the chance of false negatives, such as the classification of melanoma as a benign nevus.

$$\rho = \frac{\text{No. of malignant pixels}}{\text{No. of malignant pixels + No. of benign pixels}} \tag{3}$$

$$\hat{y}_s = \begin{cases} \text{Malanoma} & \text{if } \; \rho \geq t_r \\ \text{Benign Nevus}, & \text{Otherwise} \end{cases} \tag{4}$$

D. Evaluation Metrics

We report accuracy, F1, sensitivity, and specificity metrics. Let TP, FN, FP, and TN describe true positive, false negative, false positive, and true negative predictions, respectively. Accuracy, $(TP + TN)/(TP + FN + FP + TN)$, is the percentage of accurate predictions made by the model. $F1 = 2 \cdot (\text{precision} \cdot \text{recall})(\text{precision} + \text{recall})$, is harmonic mean where Recall = Sensitivity = $TP/(TP + FN)$ and Precision = $TP/(TP+FP)$. The sensitivity measures patches detected as malignant lesions were actual malignant class. The specificity, termed as $TN/(TN + FP)$, describes the proportion of correctly classified benign lesion patches.

E. Implementation Details

The method was implemented on the Pytorch [1] DL framework, and patch extraction was accomplished using Pyvips [2] library. VGG16 [10] feature extractor was initialized with ImageNet weights, and the classifier was initialized with random weights. We used cross-entropy loss, Adam optimizer, the learning rate of 0.0003, batch size of 256, and early-stopping of 8 epochs on validation accuracy. All models were developed on a Nvidia A100 40GB GPU. Source code is available at Github.

[1] https://github.com/pytorch/pytorch
[2] https://libvips.github.io/pyvips/

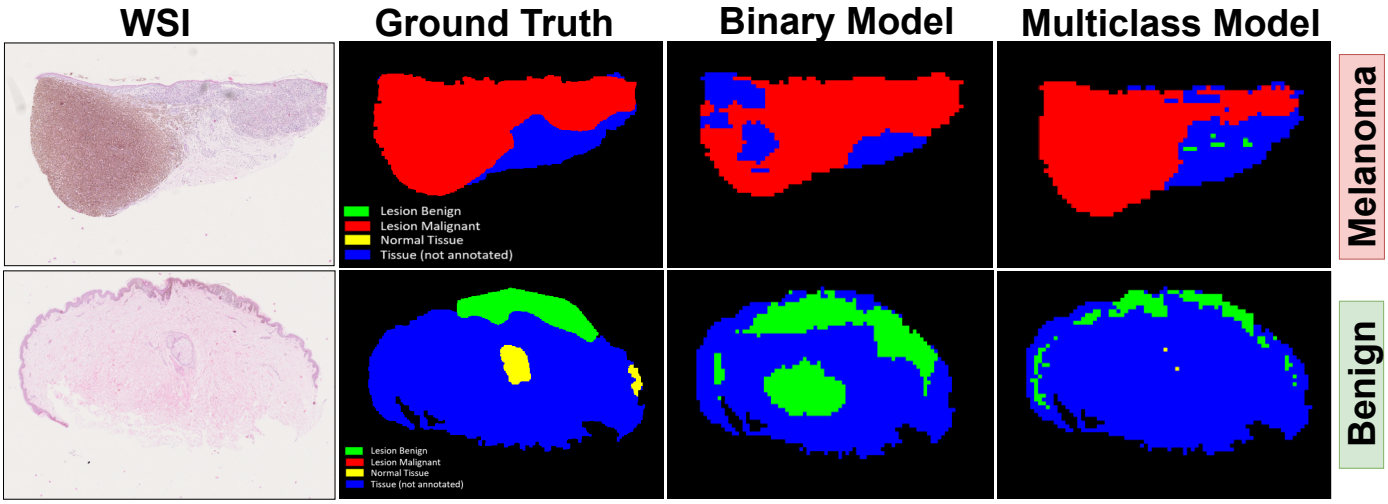

Figure 3. **Localization map for whole slide images (WSIs):** The first column represents the original WSI. The second column displays our ground truth (annotations by a pathologist). The third column shows the outcome of the binary model, and finally, the fourth column shows the outcome of the multiclass model. All WSIs, in the test set, with melanoma are correctly detected and localized.

Table I
THE RESULTS FROM THE PATCH-WISE CLASSIFICATION ON THE VALIDATION AND TEST SET.

	Models	Acc. (%)	F1	Sens.	Spec.
Validation set	Logu *et al.* [16]	96.5	0.965	0.957	0.977
	Zhang *et al.* [17]	95.5	0.976	0.986	-
	Wang *et al.* [18]	94.9	-	0.947	0.953
	$\text{Model}_{baseline}$	93.1	0.948	0.926	0.942
	$\text{Model}_{2.5x}$	**99.32**	**0.995**	0.990	**0.999**
	Model_{10x}	98.13	0.986	**0.994**	0.954
	Model_{40x}	96.02	0.971	0.986	0.906
	$\text{Model}_{multiclass}$	96.63	0.979	0.986	0.882
Test set	$\text{Model}_{2.5x}$	99.07	0.990	0.989	0.998
	$\text{Model}_{multiclass}$	98.27	0.992	0.990	0.973

IV. Results and Discussion

A. Patch-wise Classification

In this experiment, we compare the patch-wise classification performance of binary and multiclass models against the baseline model and literature. Table I shows the results of our proposed method on both the validation and test set. Binary models developed on lower magnification levels perform relatively better than other counterparts. $\text{Model}_{2.5x}$ outperforms all other models in most evaluation metrics. It significantly outperformed the baseline model with a nearly 5% F1 score. However, Model_{10x} gives the highest sensitivity value, which shows that context is more important than cellular details at a high level. $\text{Model}_{multiclass}$ resulted in higher accuracy than the baseline model and literature but did not surpass binary models. It might be due to the misclassification of normal tissue as benign lesions. In addition to this, benign and malignant lesions are located close to the epidermis and may contain some normal tissue due to the rough annotations. On the test set, multiclass models exhibited higher F1 and sensitivity in predicting malignant lesions. In contrast, binary models resulted in false positives and predicted normal adnexal structures as benign lesions.

Table II
THE RESULTS FOR THE SLIDE-LEVEL PREDICTION ON THE TEST SET, USING BEST BINARY AND MULTICLASS MODELS.

	Melanoma	Benign Nevus	Total (Acc.)
$\text{Model}_{2.5x}$	4	5	9 (100 %)
$\text{Model}_{multiclass}$	4	5	9 (100 %)
Ground Truth	4	5	

B. Slide-level Prediction

We choose the binary $\text{Model}_{2.5x}$ and $\text{Model}_{multiclass}$ from the previous experiment for slide-level prediction. The values for t_p and t_r are found to be 0.999 and 0.04, respectively, based on the validation set and fixed for the test. All patients in the test set were accurately predicted as shown in Table II.

Fig. 3 depicts the localization maps over some example WSIs from the test set. In the first row, the binary model detects some unannotated tissue as a malignant lesion. Besides, the multiclass model closely identifies melanoma, as marked in the annotation. It points to small benign lesions on unannotated areas as well. In the second row, the binary model predicts normal adnexal structure as a benign lesion because the model is unaware of normal tissue morphology. The multiclass model predicts some tissue (on the left and right) as benign lesions. Overall, the binary model overestimates the size of lesions; however, the multiclass model tightly localizes lesions inside the annotated boundary. Both models successfully predict the presence of melanoma and benign nevus at the slide level. The proposed method is computationally friendly for clinical

practice and takes roughly three minutes per WSI to obtain a diagnostic outcome and ROI.

V. Conclusion and Future Work

Computational pathology systems may assist pathologists by providing automatic diagnosis and locating regions of interest. In this work, we proposed a CNN-based method to classify patches and create a localization map to detect and segment lesions in whole slide images from skin biopsies and identify patients with melanoma. Interestingly, our method uses a single CNN network to perform lesion segmentation and do slide-level melanoma classification with one extra learned parameter. Models developed on lower magnification levels accurately provide slightly better patch-wise results indicating that context is important. Overall, the proposed method gives promising results and demonstrates its efficacy as a diagnostic service for clinical practices.

In future work, the proposed method should be validated with a larger dataset. The inclusion of tumor necrosis annotations may be beneficial as an additional diagnostic factor, as it is a vital hallmark of rapid cell proliferation.

Compliance with Ethical Standards

This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the ethics committee of the hospital (No#.2019/747/REK vest).

Acknowledgment

This work is financially supported by CLARIFY Project. CLARIFY is a research and innovation program under the Marie Skłodowska-Curie grant agreement No. 860627. The work of Helga Hardardottir is financed through the project "Pathology Services in the Western Norway Health Region – a center for applied digitization" from a Strategic investment from the Western Norway Health Authority. The authors have no relevant financial or non-financial interests to disclose.

References

[1] S. C. Chen, M. L. Pennie, P. Kolm, E. M. Warshaw, E. L. Weisberg, K. M. Brown, M. E. Ming, and W. S. Weintraub, "Diagnosing and managing cutaneous pigmented lesions: primary care physicians versus dermatologists," *Journal of general internal medicine*, vol. 21, no. 7, pp. 678–682, 2006.

[2] A. C. Society, "1.800.227.2345," in *cancer.org*. https://www.cancer.org/content/dam/CRC/PDF/Public/8823.00.pdf, January 12, 2022.

[3] N. Kanwal, S. Fuster, F. Khoraminia, T. C. Zuiverloon, C. Rong, and K. Engan, "Quantifying the effect of color processing on blood and damaged tissue detection in whole slide images," in *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*. IEEE, 2022, pp. 1–5.

[4] Z. Tabatabaei, A. Colomer, K. Engan, J. Oliver, and V. Naranjo, "Residual block convolutional auto encoder in content-based medical image retrieval," in *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*. IEEE, 2022, pp. 1–5.

[5] S. Fuster, F. Khoraminia, U. Kiraz, N. Kanwal, V. Kvikstad, T. Eftestøl, T. C. Zuiverloon, E. A. Janssen, and K. Engan, "Invasive cancerous area detection in non-muscle invasive bladder cancer whole slide images," in *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*. IEEE, 2022, pp. 1–5.

[6] N. Kanwal, E. Trygve, K. Farbod, Z. Tahlita CM, and E. Kjersti, "Vision transformers for small histological datasets learned through knowledge distillation," in *Lecture Notes in Computer Science*, vol. 13937. 27th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) Proceedings, 2023.

[7] E. Abels, L. Pantanowitz, F. Aeffner, M. D. Zarella, J. van der Laak, M. M. Bui, V. N. Vemuri, A. V. Parwani, J. Gibbs, E. Agosto-Arroyo *et al.*, "Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the digital pathology association," *The Journal of pathology*, vol. 249, no. 3, pp. 286–294, 2019.

[8] N. Wahab, I. M. Miligy, K. Dodd, H. Sahota, M. Toss, W. Lu, M. Jahanifar, M. Bilal, S. Graham, Y. Park *et al.*, "Semantic annotation for computational pathology: Multidisciplinary experience and best practice recommendations," *The Journal of Pathology: Clinical Research*, vol. 8, no. 2, pp. 116–128, 2022.

[9] N. Kanwal, F. Pérez-Bueno, A. Schmidt, K. Engan, and R. Molina, "The devil is in the details: Whole slide image acquisition and processing for artifacts detection, color variation, and data augmentation: A review," *IEEE Access*, vol. 10, pp. 58 821–58 844, 2022.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2015.

[11] M. A. Sheha, M. S. Mabrouk, A. Sharawy *et al.*, "Automatic detection of melanoma skin cancer using texture analysis," *International Journal of Computer Applications*, vol. 42, no. 20, pp. 22–26, 2012.

[12] M. H. Jafari, S. Samavi, N. Karimi, S. M. R. Soroushmehr, K. Ward, and K. Najarian, "Automatic detection of melanoma using broad extraction of features from digital images," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2016, pp. 1357–1360.

[13] S. Morales, K. Engan, and V. Naranjo, "Artificial intelligence in computational pathology–challenges and future directions," *Digital Signal Processing*, vol. 119, p. 103196, 2021.

[14] L. Tomasetti, L. J. Høllesli, K. Engan, K. D. Kurz, M. W. Kurz, and M. Khanmohammadi, "Machine learning algorithms versus thresholding to segment ischemic regions in patients with acute ischemic stroke," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 2, pp. 660–672, 2021.

[15] K. Rajitha, S. Bhat, P. Prakash, R. Rao, and K. Prasad, "Classification of microscopic images of unstained skin samples using deep learning approach," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2022, pp. 1–4.

[16] F. De Logu, F. Ugolini, V. Maio, S. Simi, A. Cossu, D. Massi, I. A. for Cancer Research (AIRC) Study Group, R. Nassini, and M. Laurino, "Recognition of cutaneous melanoma on digitized histopathological slides via artificial intelligence algorithm," *Frontiers in oncology*, vol. 10, p. 1559, 2020.

[17] D. Zhang, H. Han, S. Du, L. Zhu, J. Yang, X. Wang, L. Wang, and M. Xu, "Mpmr: Multi-scale feature and probability map for melanoma recognition," *Frontiers in Medicine*, vol. 8, p. 775587, 2022.

[18] L. Wang, L. Ding, Z. Liu, L. Sun, L. Chen, R. Jia, X. Dai, J. Cao, and J. Ye, "Automated identification of malignancy in whole-slide pathological images: identification of eyelid malignant melanoma in gigapixel pathological slides using deep learning," *British Journal of Ophthalmology*, vol. 104, no. 3, pp. 318–323, 2020.

[19] L. Tomasetti, M. Khanmohammadi, K. Engan, L. J. Høllesli, and K. D. Kurz, "Multi-input segmentation of damaged brain in acute ischemic stroke patients using slow fusion with skip connection," *arXiv preprint arXiv:2203.10039*, 2022.

[20] R. Wetteland, K. Engan, and T. Eftesol, "Parameterized extraction of tiles in multilevel gigapixel images," in *2021 12th International Symposium on Image and Signal Processing and Analysis (ISPA)*. IEEE, 2021, pp. 78–83.