A Dual Convolutional Neural Network Pipeline for Melanoma Diagnostics and Prognostics

Marie Bø-Sande[1,†], Edvin Benjaminsen[1,†], Neel Kanwal[1], Saul Fuster[1], Helga Hardardottir[2,3], Ingrid Lundal[2,3], Emiel A.M. Janssen[2,3], and Kjersti Engan[*1]

[1] Department of Electrical Engineering and Computer Science, University of Stavanger, Stavanger 4021, Norway
[2] Department of Chemistry, Bioscience and Environmental Engineering, University of Stavanger, Stavanger 4021, Norway
[3] Department of Pathology, Stavanger University Hospital, Stavanger 4011, Norway
[†]These authors contributed equally.
{neel.kanwal, saul.fusternavarro, kjersti.engan}@uis.no

1 Abstract

Melanoma is a type of cancer that begins in the cells controlling the pigment of the skin, and it is often referred to as the most dangerous skin cancer. Diagnosing melanoma can be time-consuming, and a recent increase in melanoma incidents indicates a growing demand for a more efficient diagnostic process. This paper presents a pipeline for melanoma diagnostics, leveraging two convolutional neural networks, a diagnosis, and a prognosis model. The diagnostic model is responsible for localizing malignant patches across whole slide images and delivering a patient-level diagnosis as malignant or benign. Further, the prognosis model utilizes the diagnostic model's output to provide a patient-level prognosis as good or bad. The full pipeline has an F1 score of 0.79 when tested on data from the same distribution as it was trained on.

2 Introduction

Melanoma cancer is the leading cause of death from skin disease. It begins in the skin cells called melanocytes, and around 30% starts in existing moles. According to a recent worldwide study, the number of newly diagnosed melanoma cases will rise by more than 50%, up to 510,000 by 2040, while the number of melanoma deaths will rise by almost 68%, from 57,000 in 2020 to 96,000 in 2040 [1]. In regards to Norway, the annual cancer report shows a 20% increase in melanoma cases from 2021 to 2022 [2]. Coupled with the increased incidence rate, the estimated survival rate for five years following diagnosis varies depending on the stage of melanoma [3]. Consequently, early detection of melanoma plays a crucial role in the prognostic outcome.

In recent years, digital pathology is becoming mainstream, producing whole slide images (WSIs) as digital microscopy gigapixel images of tissue slides. The process of producing WSIs from biopsies, with its potential artifacts, is described in [4]. Computational pathology (CPATH) is a growing field dealing with automated solutions for visualization, diagnostics, and prognostics from WSI using image processing and deep learning (DL).

Melanoma detection using DL techniques has shown promising results, which can help with early diagnosis and treatment decisions [5–7]. DL algorithms can detect potential regions with melanoma by identifying various cellular and tissue-level features and enhancing diagnostic accuracy [8]. Some existing methods focus on detecting melanoma by classifying tissue samples and moles either as melanoma or benign nevi [5, 6, 9]. Others address the question of prognostic prediction from the lesion-tissue of verified melanoma cases, typically with the lesion manually delineated [10–12]. Clinical labels are usually patient-based, and providing manually annotated regions for the melanoma tissue is time-consuming. The challenge of having a complete CPATH pipeline that can differentiate between WSIs with benign nevi and melanoma (malignant), segment the melanoma region from a WSI, and make prognostic predictions in the case of melanoma is currently unexplored.

To address this challenge in this work, we are developing a pipeline by integrating two convolutional neural networks (CNN), as illustrated in Figure 1. By leveraging DL methodologies, the first model of the pipeline identifies melanoma in WSIs and produces a patient-level diagnosis, whereas the second model provides a prognosis for identified melanoma patients. The performance of each CNN model is evaluated individually before integrating them into a comprehensive pipeline.

3 Data Materials

The dataset is collected at Stavanger University Hospital (SUH), Stavanger, Norway. A Hamamatsu Nanozoomer s60 scanner was used to scan a cohort of Hematoxylin and Eosin (H&E) stained glass slides

*Corresponding Author

Proceedings of the 5th Northern Lights Deep Learning Conference (NLDL), PMLR 233, 2024.

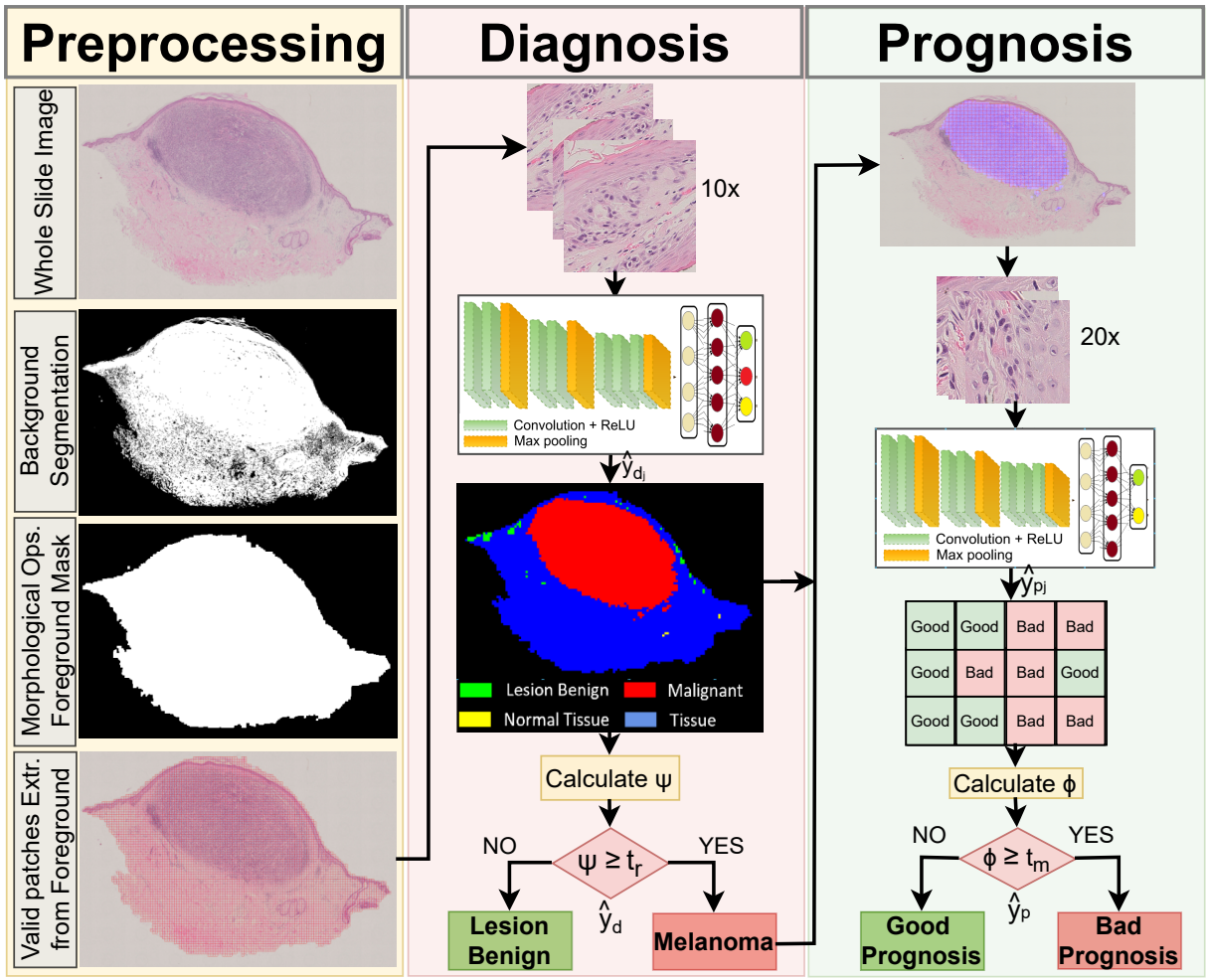

Figure 1. An overview of our proposed deep learning pipeline. *Preprocessing:* Background/foreground segmentation is applied on the whole slide image (WSI) to find tissue regions. Later, morphological operations remove small holes and separate small areas, and then patches are extracted at 10x magnification for the diagnosis model. *Diagnosis:* Prediction for every patch is used to calculate patient-level prediction as benign vs. malignant. *Prognosis:* Detected lesion regions on WSIs predicted as malignant are used to evaluate prognosis at the patient level as good or bad.

at 40x magnification, thereafter saved in NDPI format. Clinical labels give the patient diagnosis as $y_i^d \in \{0, 1\}$ where 1 indicates melanoma and 0 benign nevi, and i is a patient (or WSI) index, and d indicates diagnostic label. For a patient diagnosed with melanoma, the prognostic label was established using follow-up data by considering the occurrence of either local or distant metastasis (bad), or the absence of metastasis (good), within a five-year time frame; $y_i^p \in \{0, 1\}$ where 1 indicates bad prognosis and 0 good prognosis. The dataset is divided into annotated WSIs, D^a, and not-annotated WSIs, $D^{n/a}$, explained more in the following subsections. Later, patches were extracted from the WSIs for analysis. A breakdown of the number of patches extracted from the different sets is shown in Tab. 2.

Table 1. The amount of patches from each subset of D_d^a after patch extraction for the diagnostic model.

Label	D_{train}	D_{val}	D_{test}	Total
B	42 420	11 660	7 190	61 270
M	215 320	25 169	37 310	27 799
NE	2 801	657	604	4 062
Total	260 541	37 486	45 104	343 131

Table 2. The amount of patches of D_p^a after patch extraction for the prognostic model.

Label	Total
Good	5 542
Bad	6 358
Total	11 900

3.1 Annotated WSIs

D^a is a set of 125 WSIs, 47 benign nevi, and 78 with melanoma. The lesion, or region of interest (ROI), in all WSI, is roughly annotated by a pathologist. The annotated regions of the lesion have two different classes, corresponding to melanoma M and benign nevi B. In addition, some areas of normal epidermal tissue NE are annotated, but not all such areas. Tissue outside these regions is *not* annotated. Labels associated with these regions are defined as y^d_{ji}, where j is a patch index, i is a patient or WSI index. In addition, there are large tissue regions that are *not* annotated in all WSI. The diagnosis model used a sub-dataset D^a_d of 90 WSIs, where 73 of them were used for training, 8 for validating, and 9 for testing, with approximately 50% benign nevi and melanoma. The prognostic model used a sub-dataset D^a_p of 52 WSIs, all with melanoma, 50% with bad, and 50% with good prognosis. Some patients were excluded as they were present in both D^a_d and D^a_p. A total of 9 bad prognosis and 5 with good prognosis. A sub dataset $\hat{D}^a_p$ was defined comprising all the images from D^a_p except for the aforementioned excluded patients, as they were employed for the development of the diagnostic model. There is no overlap from training to validation or test in the pipeline experiments we show.

3.2 Non-annotated WSIs

A dataset $D^{n/a}$ containing 243 WSIs from the SUH cohort is provided with patient-level clinical labels without any manual annotation around lesion areas. Of all 243 WSIs, 110 of them were diagnosed with a benign nevis, and 133 with melanoma; 10 of these had bad prognosis (metastasis within five years). The dataset is divided into a train/validation set of 203 WSIs and a test set of 40 WSIs. In the test set, 18 WSIs are labeled as benign and 22 as melanoma, 2 of them having a bad prognosis (metastasis within five years).

4 Method

4.1 Preprocessing

To enable the analysis of WSIs using CNN, the tissue regions within the WSIs were divided into smaller patches, of size 256×256 pixels at different magnification levels (2.5x, 10x, and 40x). Let $\mathbf{x}^{10x}_{ji}$ denote patch j from WSI i at magnification level 10x. The index i denotes the WSI and is sometimes omitted. To separate the tissue from the background, background-foreground segmentation was performed by transforming the RGB images to the HSV color space, and the Hue channel was thresholded within the range of [100-180] to identify purple and pink tones. Morphological opening and closing operations were applied to close holes in the foreground and remove small areas. Grid extraction was applied to extract valid patches as described in [13].

4.2 Diagnosis

Valid patches, $\mathbf{x}^{10x}_{ji}$, from the preprocessing of WSI i are fed into the diagnosis model, providing a patch-level prediction: $f^d(\mathbf{x}^{10x}_{ji}) = \hat{y}^d_{ji}$. The feature extractor of the model is based on the VGG16 [14] architecture with pre-trained weights from ImageNet [15]. A three-layer classifier of fully connected layers is added. The diagnostic model is fine-tuned using the annotated training data from D^a_d, as in [6]. The models output layer consists of a softmax giving an array $\mathbf{v}_{ji}$ of three probability values for each patch, benign nevi (B), melanoma ($\mathbf{M}$), and normal epidermal tissue (NE). The patch-level diagnosis predictions are denoted as $\hat{y}^d_{ji}$, for patch j, and WSI i. If $max(\mathbf{v}_{ji}) > t_p$, $\hat{y}^d_{ji}$ is set to the most probable class label (M, B, NE). Else, $\hat{y}^d_i$ is set to T for tissue (i.e., none of the other classes). Thus, we train the model with three labels, but we classify the patches into four classes, $\hat{y}^d_{ji} \in \{M,B,NE,T\}$. The patient-level prediction $\hat{y}_i$ is determined by calculating the ratio ψ of number of patches predicted as malignant (i.e. melanoma) over other patches. In this work, two different methods are used to calculate the ratio as shown in Eq. (1) and (2). The first ratio, ψ^{MB}, calculates the ratio between predicted malignant and benign patches, while the second ratio, ψ^{MT} calculates the ratio between malignant patches and patches in the entire tissue mask. Let the indicator function, $I(\hat{y}_{ji}, \{M,B\}) = 1$ if $\hat{y}_{ji} = M$ or B and 0 otherwise:

$$\psi^{MB}_i = \frac{\sum_j I(\hat{y}^D_{ji}, M)}{\sum_j I(\hat{y}^D_{ji}, \{M,B\})} \tag{1}$$

$$\psi^{MT}_i = \frac{\sum_j I(\hat{y}^D_{ji}, M)}{\sum_j I(\hat{y}^D_{ji}, \{M,B,NE,T\})} \tag{2}$$

Thereafter, the ratio is compared with a threshold t_r. If $\psi_i < t_r$ the WSI i is predicted as benign (0); conversely, if $\psi_i > t_r$ the WSI i is predicted as melanoma (1), i.e. finding patient-level diagnosis label $\hat{y}^d_i$ as shown in Eq. (3).

$$\hat{y}^d_i = \begin{cases} 1, (melanoma) & \text{if } \psi_i \geq t_r \\ 0, (benign) & \text{else} \end{cases} \tag{3}$$

Let $\{\mathbf{x}_j\}_M$ denote the set of patches where $I(\hat{y}^d_j, M) \cap \hat{y}^d$, i.e. when the patch is predicted as malignant and the patient is predicted as melanoma, defines the ROI_d for further prognostic analysis.

4.3 Prognosis

The prognostic model utilizes a VGG16 backbone, and transfer learning is used with pretrained weights.

The classifier of the VGG16 is replaced with three fully-connected layers, the last having softmax activation function, giving a binary output for good or bad prognosis, trained on D_p^a as in [12]. The prognosis model uses malignant patches of the predicted melanoma WSIs for further analysis, i.e. the ROI_d defined by $\{\mathbf{x}_j\}_M$ as described in the previous section. However, the prognostic model operates on a different magnification scale than the diagnostic model. Thus, the ROI from the malignant patches is used to extract new patches at 20x magnification with the method of [13], requiring minimum 70% overlap between a valid patch and the ROI.

The prognosis model provides a patch-level prediction: $f^p(\mathbf{x}_{ki}^{20x}) = \hat{y}_{ki}^p$ for image i and patch $k \in \text{ROI}_d$ defined by $\{\mathbf{x}_j\}_{Mi}$. $\hat{y}_{ki}^p \in \{0, 1\}$ where "1" indicates bad prognosis at patch level. A threshold t_m is used to calculate patient-level $\hat{y}^p$ with "1" indicating a bad prognosis and "0" indicating a good prognosis at the WSI level, as shown in Eq. (4).and Eq. (5).

$$\phi_i = \frac{\sum_j \hat{y}_{ji}^p}{\sum_j (I(\hat{y}_{ji}^p, 0) + I(\hat{y}_{ji}^p, 1))} \tag{4}$$

$$\hat{y}_i^p = \begin{cases} 1, & \text{if } \phi_i \geq t_m \\ 0, & \text{else} \end{cases} \tag{5}$$

4.4 Evaluation Metrics

Let TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative, respectively. The accuracy is calculated by $(TP+TN)/(TP+FP+FN+TN)$. Recall/ sensitivity is calculated using $TP/(TP+FN)$. Precision $= TP/(TP+DP)$ and specificity $= TN/TN+FP$. F1 is a weighted harmonic mean of precision and recall.

5 Experiments and Results

Experiments are done by validating the models on the annotated datasets D^a. The pipeline is tested by comparing prognosis prediction using annotated melanoma ROI as inputs with prognosis prediction using automated ROI, i.e., outputs from the diagnostic model, on the exact same dataset. The performance of both models is tested with dataset $D^{n/a}$. We employed 5-fold cross-validation for training our models. We selected the best-performing model and proceeded with it. The results we have shown correspond to the values of the best-performing model trained using cross-validation. By evaluating the performance of each model separately, it is possible to assess their individual accuracy in addition to testing the combined pipeline.

5.1 Annotated data

The diagnostic model achieved a performance of 100% accuracy at the WSI level on the 9 WSIs in the test set of dataset D_d^a.

5.1.1 Full pipeline test

The prognosis model's validation process involves comparing the performance of f^p when running on a dataset $\hat{D}_p^a$ with ROI inputs from the manual annotations, ROI_a, and with ROI from the masks generated by the predictions from the diagnosis model f^d, ROI_d. By utilizing the diagnosis model's outputs as inputs for the prognostic model's evaluation, we can effectively assess the performance and accuracy of the prognostic model in a realistic setting. This approach allows us to better understand how the two models work together and how well the prognostic model performs when applied to new, unseen data.

Table 3 displays the evaluation metrics after running the prognosis on the dataset ($\hat{D}_p^a$).

Table 3. Evaluation Metrics after running the prognosis on dataset $\hat{D}_p^a$ with annotation masks and generated masks from the diagnosis.

$\hat{D}_p^a$	Sens.	Spec.	F1	Accuracy
ROI_a	0.941	**0.714**	**0.821**	**0.816**
ROI_d	**1.000**	0.571	0.791	0.763

The F1 score and Accuracy are slightly better when using ROI_a from manual annotations, but the results look promising for using the automatically found ROI_d.

5.2 Non-annotated data

In this experiment, the diagnostic model's performance on the non-annotated dataset $D_{val}^{n/a}$ will be evaluated. The initial thresholds for patch-level classification (t_p) and patient-level classification (t_r) are set at 0.999 and 0.04, respectively. These thresholds were found in the previous experiment, where the model predicted all images correctly on D_d^a. The results of the experiment can be found in Table 4.

Table 4. Results from running inference with diagnosis model on $D_{val}^{n/a}$ with thresholds $t_p = 0.999$ and $t_r = 0.04$.

Eval. Metric	Sens.	Spec.	F1	Accuracy
Score	0.977	0.146	0.728	0.601

The diagnostic model's accuracy is measured to be 0.601, indicating that it correctly classifies only 60% of the 243 WSIs in dataset $D^{n/a}$. Furthermore, the recall value is calculated to be 0.977, indicating that the model excels at correctly predicting almost all melanoma cases but faces difficulties in accurately predicting benign cases.

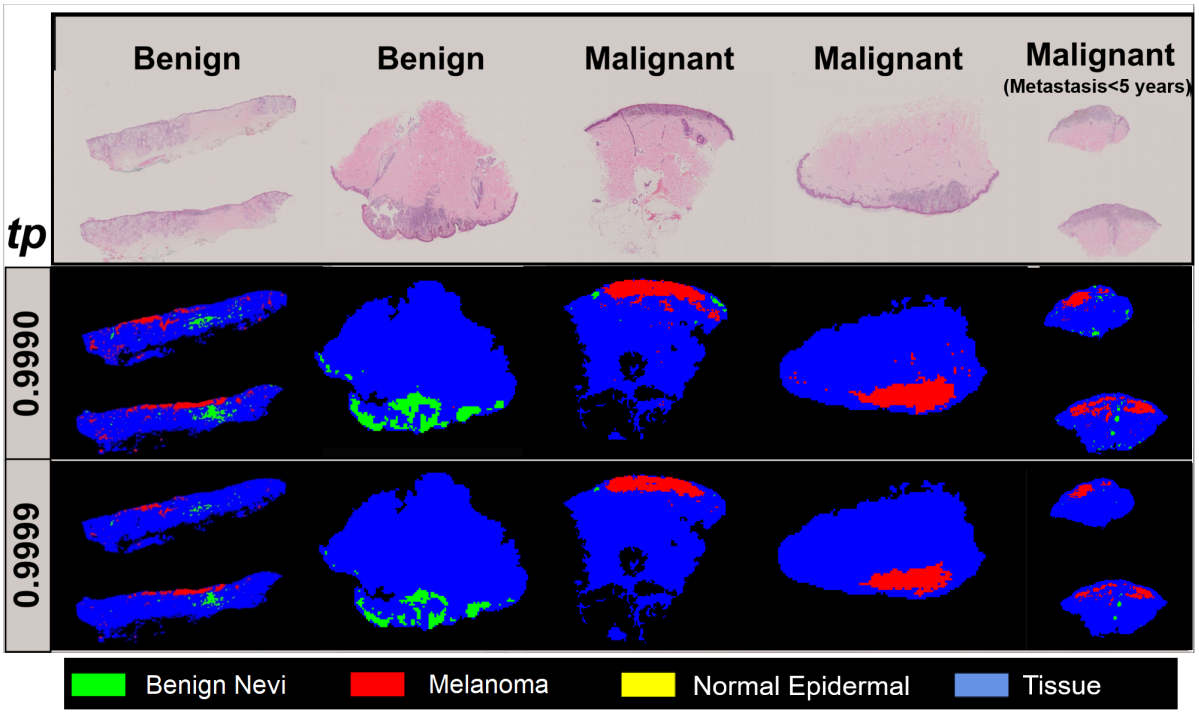

Figure 2. Examples of the diagnosis model's patch predictions at different t_p thresholds for $D_{val}^{n/a}$. Each column displays a WSI with its ground truth label (patient level) at the top. Each row presents the prediction masks for each WSI with the given t_p on the left side.

Table 5. Non-annotated data, pipeline test on the diagnosis (f^d) and prognostic model (f^p) on the $D_{test}^{n/a}$. For each of the three tests, performance metrics for the diagnostic model only (D), as well as the complete pipeline (P), are reported as D/P in the metrics.

Test	Mod	t_p	t_r	ψ	Spec.	Sens.	F_1	Acc
1	D/P	0.9990	0.04	MB	0.11/0.14	**1.00**/1.00	**0.73**/0.11	0.60/0.18
2	D/P	0.9990	0.04	MT	**0.39**/0.07	0.82/1.00	0.71/0.14	0.63/0.14
3	D/P	**0.9999**	**0.01**	**MT**	**0.39**/0.11	0.86/1.00	0.73/0.14	**0.65**/0.17

The results from this evaluation demonstrate that the diagnosis model shows some level of generalizability for new data with the current settings. However, it is evident that there is room for improvement to enhance its performance further. This highlights the need to focus on parameter tuning and optimization for the model.

Dataset $D_{val}^{n/a}$ is used to find optimal threshold t_p and t_r. Figure 2 shows predictions based on different t_p. The first column shows an example of a benign slide that is mistaken for melanoma. There are very few predictions of the *NE* class, probably because the class is underrepresented during training. The distinction between *NE* and *T* has no diagnostic or prognostic relevance, but we keep it as a separate class in case it helps separate benign from malignant cases. In future work, we will investigate this further.

When a lesion is annotated, it is obvious to calculate the ratio of MB patches within the lesion. However, in situations where the lesion size, as well as the tissue size, varies and is unknown, focusing on MT appears to be a more logical approach. For each t_p, different t_r was tested with a ROC plot to find the optimal threshold, Figure 3 shows the MT ROC plot for $t_p = 0.9999$ and $t_r = 0.01$ marked

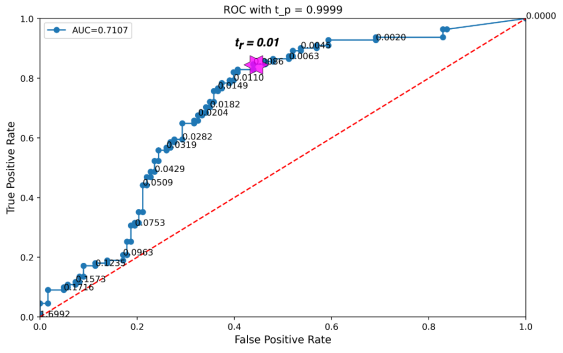

Figure 3. ROC plot for diagnosis model with $t_p = 0.9999$ and MT_{rate} method on dataset $D^{n/a}$.

as updated thresholds for the new, larger dataset. For MB, we observed that the optimal threshold choice yields relatively poorer performance across all t_r values. All in all, the increase in AUC score suggests that MT is more effective.

5.2.1 Non-annotated data pipeline test

A final test is done using the updated thresholds from the previous experiment, and testing the entire pipeline on the test set from $D^{n/a}$. Results are presented in Table 5. The diagnose model performs reasonably well, with F1 scores around 0.73. The model recall is higher than the specificity, which is also what is desired since it is better to be sure that malignant melanoma is discovered. The difference between the new and old thresholds is not very large. The prognostic model, however, has a recall of 1 in all experiments and a bad specificity, even if the results in Table3 are promising. Prognosis prediction is far more difficult than diagnostic prediction in general, and this shows us that the model has not generalized well enough, and a larger training set with both good and bad prognoses is needed to get general models. In future work, we will investigate using multiple instance learning with a larger dataset for the prognostic part.

6 Conclusion

This paper presents a pipeline putting together two CNN models, one for melanoma detection and localization and one for prognosis prediction on melanoma cases. The pipeline test demonstrates that the prognostic model works similarly well, with F1 scores of 0.82 and 0.79 if the input to the model comes from manual annotation or the output of the diagnostic model when tested on the data set from the same distribution as used to train the models, which is very encouraging. Further updating of the parameters in the diagnostic model showed reasonably good performance for the diagnosis part on a new data set; however, the prognostic model overestimates bad prognosis and should be trained on larger data sets. Prognosis is generally harder to predict.

Compliance with ethical standards

This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Regional Ethics Committee (No: 2019/747/RekVest). The authors have no relevant financial or non-financial interests to disclose.

Acknowledgements

Thanks to Andres David Mosquera Zamudio for verifying some of the slides.
This research has received funding from the European Union's Horizon 2020 research and innovation program under grant agreements 860627 (CLARIFY) and "Pathology Services in the Western Norway Health Region – a centre for applied digitization (PiV)" from a Strategic investment from the Western Norway Health Authority.

References

[1] M. Arnold, D. Singh, M. Laversanne, J. Vignat, S. Vaccarella, F. Meheus, A. E. Cust, E. de Vries, D. C. Whiteman, and F. Bray. "Global burden of cutaneous melanoma in 2020 and projections to 2040". In: *JAMA dermatology* 158.5 (2022). doi:10.1001/jamadermatol.2022.0160, pp. 495–503.

[2] C. R. of Norway. *Cancer in Norway 2022 - Cancer incidence, mortality, survival and prevalence in Norway*. Oslo: Cancer Registry of Norway, 2022.

[3] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal. "Cancer statistics, 2023". In: *Ca Cancer J Clin* 73.1 (2023). doi:10.3322/caac.21763, pp. 17–48.

[4] 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". In: *IEEE Access* 10 (2022). 10.1109/ACCESS.2022.3176091, pp. 58821–58844.

[5] L. Launet, A. Colomer, A. Mosquera-Zamudio, A. Moscardó, C. Monteagudo, and V. Naranjo. "A Self-Training Weakly-Supervised Framework for Pathologist-Like Histopathological Image Analysis". In: *2022 IEEE International Conference on Image Processing (ICIP)* (2022). doi:10.1109/ICIP46576.2022.9897274, pp. 3401–3405.

[6] N. Kanwal, R. Amundsen, H. Hardardottir, L. Tomasetti, E. Sand, E. A. Janssen, K. Engan, et al. "Detection and localization of melanoma skin cancer in histopathological whole slide images". In: *2023 31st European Signal Processing Conference (EUSIPCO)*. IEEE. 2023, pp. 975–979.

[7] D. Sauter, G. Lodde, F. Nensa, D. Schadendorf, E. Livingstone, and M. Kukuk. "Deep learning in computational dermatopathology of melanoma: A technical systematic literature review". In: *Computers in Biology and Medicine* (2023). doi:10.1016/j.compbiomed.2023.107083, p. 107083.

[8] K. Das, C. J. Cockerell, A. Patil, P. Pietkiewicz, M. Giulini, S. Grabbe, and M. Goldust. *Machine learning and its application in skin cancer*. doi:10.3390/ijerph182413409. 2021.

[9] R. del Amor, F. J. Curieses, L. Launet, A. Colomer, A. Moscardó, A. Mosquera-Zamudio, C. Monteagudo, and V. Naranjo. "Multi-Resolution Framework For Spitzoid Neoplasm Classification Using Histological Data". In: *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)* (2022). doi:10.1109/IVMSP54334.2022.9816260, pp. 1–5.

[10] J. Hu, C. Cui, W. Yang, L. Huang, R. Yu, S. Liu, and Y. Kong. "Using deep learning to predict anti-PD-1 response in melanoma and lung cancer patients from histopathology images". In: *Translational Oncology* 14 (2020). doi:10.1016/j.tranon.2020.100921.

[11] S. Forchhammer, A. Abu-Ghazaleh, G. Metzler, C. Garbe, and T. Eigentler. "Development of an image analysis-based prognosis score using google's teachable machine in Melanoma". In: *Cancers* 14.9 (2022). doi:10.3390/cancers14092243, p. 2243.

[12] C. Andreassen, S. Fuster, H. Hardardottir, E. A. M. Janssen, and K. Engan. "Deep Learning for Predicting Metastasis on Melanoma WSIs". In: *ArXiv* abs/2303.05752 (2023). doi:10.48550/arXiv.2303.05752.

[13] 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)*. doi:10.1109/ISPA52656.2021.9552104. 2021, pp. 78–83.

[14] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014). doi:10.48550/arXiv.1409.1556.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. doi:10.1109/CVPR.2009.5206848. 2009, pp. 248–255.