Contrastive and attention-based multiple instance learning for the prediction of sentinel lymph node status from histopathologies of primary melanoma tumours. *

Carlos Hernandez Perez[1 [0000−0002−0054−8573]], Marc Combalia Escudero[2,3 [0000−0001−5237−4256]], Susana Puig[2,3 [0000−0003−1337−9745]], Josep Malvehy[2,3 [0000−0002−6998−914X]], and Veronica Vilaplana Besler[1 [0000−0001−6924−9961]]

[1] Image Processing Group, Universitat Politècnica de Catalunya, Barcelona, Spain
{carlos.hernandez.p, veronica.vilaplana}@upc.edu
[2] Dermatology Department, Melanoma Unit, Hospital Clínic de Barcelona, IDIBAPS, University of Barcelona, Barcelona, Spain
[3] Athena Tech, S.L., Barcelona, Spain

Abstract. Sentinel lymph node status is a crucial prognosis factor for melanomas; nonetheless, the invasive surgery required to obtain it always puts the patient at risk. In this study, we develop a Deep Learning-based approach to predict lymph node metastasis from Whole Slide Images of primary tumours. Albeit very informative, these images come with complexities that hamper their use in machine learning applications, namely their large size and limited datasets. We propose a pre-training strategy based on self-supervised contrastive learning to extract better image feature representations and an attention-based Multiple Instance Learning approach to enhance the model's performance. With this work, we quantitatively demonstrate that combining both methods improves various classification metrics and qualitatively show that contrastive learning encourages the network to output higher attention scores to tumour tissue and lower scores to image artifacts.

Keywords: Whole Slide Image · Contrastive learning · Attention-based Multiple Instance Learning · Early detection.

1 Introduction

Skin cancer incidence has increased in the last decade worldwide [1], and it's the fourth leading cause of cancer-related mortality [16]. Melanoma constitutes the leading cause of death due to skin cancer when considering its incidence and mortality. Therefore, there is a need for early detection of melanoma: the

* Work supported by the Spanish Research Agency (AEI) under project PID2020-116907RB-I00 of the call MCIN/ AEI /10.13039/501100011033 and the project 718/C/2019 funded by Fundació la Marato de TV3.

five-year survival rate of melanoma rapidly decreases as its stage advances, from 97% for stage I to only 15-20% for stage IV.

The result of a sentinel lymph node biopsy (SLNB) is a key early prognosis factor for melanoma patients. An SLNB is a procedure in which the sentinel lymph node is identified, removed, and examined to determine whether cancer cells are present. Only patients that have already been diagnosed with cancer undergo this surgery. A negative SLNB result suggests that cancer has not yet spread to nearby lymph nodes or other organs. However, 80% of patients do not benefit from this intervention because they have unaffected lymph nodes [7]. Thus, it would be desirable to determine the likelihood of a positive SLNB in advance to reduce the number of unnecessary surgeries and their associated morbidities.

Artificial intelligence and, especially, Convolutional Neural Networks (CNN) [8] have proved to be state-of-the-art methods for medical imagining analysis [15, 18–20]. In recent years, the value of deep learning empowered computer-assisted diagnosis has been established in dermatological imaging-based decision-making models [5]. It has been successfully implemented in a variety of use cases, from finding skin lesion biomarkers [6] to identifying tumour tissue in Whole Slide Images (WSI) [13]. Unfortunately, the literature on the prediction of SLNB positivity from WSI of primary cutaneous melanoma tumours is scarce. To our knowledge, the only work that addresses this problem is Brinker et al. [3] where they use a Multiple Instance Learning (MIL) framework combining clinical data, cell features, and the slide feature vectors with a squeeze and excitation block before classification. Our approach tackles the same problem by combining self-supervised contrastive learning [11] to improve the quality of the extracted features and an attention-based MIL strategy to improve the model's classification capabilities. Additionally, the attention mechanism allows the visualization of the relative patch importance for the final model's prediction.

2 Materials and methods

In this section, we first present the dataset and briefly discuss the usage of MIL in our work. Afterwards, we describe the details of our proposed model, and, finally, we present a primer on contrastive learning and how we apply it to WSI analysis.

2.1 Dataset

Ethics approval was obtained from the ethics committee of Hospital Clinic de Barcelona before the study was initiated. A total of 195 digitised WSIs were used, each coming from a different patient. WSIs were obtained from a cutaneous biopsy of a primary melanoma tumour, while the target variable was obtained after analyzing an SLNB (107 SN negative and 88 SN positive). Both biopsies were performed, at most, within a year. Available clinical data included the patient's age, tumour thickness, and ulceration, presented in Table 1.

Table 1. Clinical data. Mean and standard deviation computed for the age and tumour thickness.

Clinical variable	SLN+	SLN-	All
Age (years)	54.2 ± 14.6	58.1 ± 14.6	56.3 ± 14.7
Ulceration (yes/no)	43/45	64/43	107/88
Tumour thickness (mm)	1.12 ± 0.69	1.26 ± 0.65	1.2 ± 0.68
SLN status	88	107	195

Patch extraction: The tissue was segmented from the background through Otsu's thresholding method [14]. Afterwards, patches of 256x256 pixels were cropped without any overlap. Then, patches containing a small tissue area were regarded as non-informative and discarded from the dataset. The average number of extracted patches per slide was 488.

2.2 Multiple instance learning

Since only labels at the WSI level are available (the positivity of the SLNB), we used a weakly supervised architecture based on Multiple Instance Learning. MIL is a family of weakly-supervised learning algorithms in which the learner receives a set of labeled bags where each contains an unconstrained number of unlabeled instances. In our case, bags are slides, S, and instances are patches, $\mathcal{X}$. An assumption needs to be made regarding the relationship between a bag, its instances, and the class label of the bag. Two main assumptions are tested in this work:

The standard assumption: each patch $x \in \mathcal{X}$ has an associated label $y \in \{0, 1\}$ which is hidden to the learner. If at least one of them has a positive label, the whole slide is considered to be positive. Formally, let $S = \{(x_1, y_1), \ldots, (x_K, y_K)\}$ be a slide represented as a set of patches with their labels. The label of S is then $c(S) = 1 - \prod_{i=1}^{K}(1 - y_k)$.

Attention-based MIL pooling: In contrast to the previous assumption, attention mechanisms allow for a more flexible and adaptive MIL pooling [10] through their ability to adjust to a task and data by having trainable parameters. Let $H = \{h_1, ..., h_K\}$ be a slide represented with a set of K patch feature embeddings, being $h_i = e(x_i)$ where e is any encoder; then a MIL pooling can be expressed as:

$$\mathbf{r} = \sum_{k=1}^{K} a_k \mathbf{h_k}; \qquad \mathbf{a} = \{a_1, ..., a_K\}; \qquad \sum_{k=1}^{K} a_k = 1 \tag{1}$$

Where a_k is the attention weight for the *k-th* patch and $\mathbf{r}$ is the aggregated patch features for the slide. Let The label of S is $c(S) = m(\mathbf{r})$, where $m(\cdot)$ is typically a multilayer perceptron (MLP).

2.3 Proposed model

Our proposed method expands the work by Ming Y. Lu et al. [13] and consists of two steps: a feature extractor and a classification module, as shown in Figure 1. The feature extractor uses the convolutional layers of a ResNet50 model to extract features from the patches, reducing the dimensionality of the data. This, in turn, enables the classification module to fit all the slide's info into a single commercial GPU. During classification, for both training and inference, the model examines and ranks all patches in the tissue regions of a WSI, assigning an attention score to each patch, which informs its contribution or importance to the collective slide-level representation for each class. Attention scores are obtained with a gated attention mechanism [10]. Next, each feature vector is forwarded through its respective class branch consisting of an MLP with a single neuron as an output. Finally, both branch outputs are forwarded together through a softmax layer. The maximum of the two values is taken as the prediction for the WSI.

Clinical data is added through an additional input head with two linear layers and non-linearities. The resulting vector and the aggregated patch features are normalized before concatenation. The combined feature vector is then forwarded through each class branch.

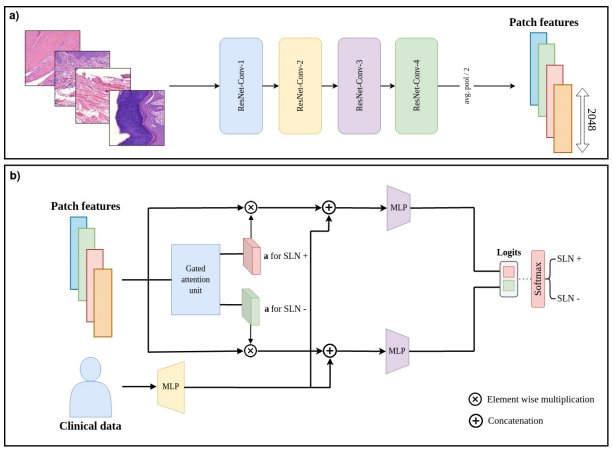

Fig. 1. Model's feature extractor a) and classification module b).

Patch importance visualization: The extracted patches' relative importance can be computed by transforming the attention scores for the model's predicted

class into percentiles and mapping them to their corresponding spatial location in the WSI. These scores are the output of the gated attention unit shown in Figure 1. The result is presented in the form of an overlapping heatmap.

2.4 Self-supervised contrastive learning:

Contrastive learning schemes have being applied in previous skin cancer related studies to learn effective visual representations without the need for data labels [2,12,17]. In this work, we use Google's Simple framework for Contrastive Learning of visual Representations (SimCLR)[4]. This framework is composed of four major components:

- From a batch of size N, *data augmentation* is performed twice to generate, from each patch x, two patches $\tilde{x}_i$ and $\tilde{x}_j$. The result is a set of 2N patches with correlated views of the same examples. The data augmentations used are: *vertical and horizontal flip* with 50% probability each, *random cropping* with a resizing of the resulting image to the original size. *Color distortions* such as small changes in the image's hue or conversion to grayscale are applied with a probability of 80% and 20%, respectively. At the end of the data augmentation pipeline, all channel values were normalized using ImageNet's mean and variance.
- A *base encoder* $f(\cdot)$, the convolutional layers of a CNN, used to extract feature vectors h_i from each cropped patch $h_i = f(\tilde{x}_i)$.
- A *small projection head* $g(\cdot)$ in the form of an MLP with one hidden layer which maps h_i to z_i the space where contrastive loss is applied. That is, $z_i = g(h_i)$ where $z_i \in R^j$, and $h_i \in R^d$, and $j < d$.
- A *contrastive loss function* that encourages the clustering of z_i and z_j from the same original image together while separating them from the rest of the 2(N-1) patches.

The loss function is:

$$L_{i,j} = -\log \frac{\exp\left(sim\left(\boldsymbol{z}_i, \boldsymbol{z}_j\right)/\tau\right)}{\sum_{k=1}^{2N} 1_{[k \neq i]} \exp\left(sim\left(\boldsymbol{z}_i, \boldsymbol{z}_k\right)/\tau\right)} \tag{2}$$

where $sim(z_i, z_j) = z_i^T z_j / ||z_i|| ||z_j||$, measures the similarity between z_i and z_j. $1_{[k \neq i]}$ is 1 if $k \neq i$, and 0 otherwise. τ denotes a temperature parameter. The total loss is computed across all pairs in a mini batch. A visual example can be found in Figure 2, for a more detailed explanation, refer to [4].

3 Experimental set-up and results

This section explains the two different sets of WSI features used and the experiments performed with them. Results are shown in Table 2 and Figure 3.

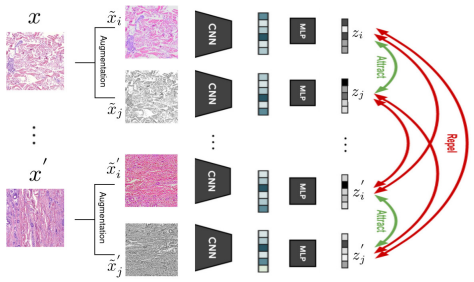

Fig. 2. Self-supervised contrastive learning framework. Each original patch x, is augmented twice and forwarded through a CNN. The features obtained after the projection head are compared using the cosine similarity. This loss function reinforces the clustering of vectors z_i and z_j while penalizing the similarity with the other 2(N-1) inputs.

3.1 Feature extraction

Two ResNet50 [9] architectures were used to extract features from the cropped patches. Their pre-training scheme differed: the first CNN was pre-trained on ImageNet while the second was additionally fine-tuned using the SimCLR framework with a batch size of 64 patches. We denote the first set of features F_{Im} and the second F_{CLR}. The dimension of each embeddding is $h_k \in R^{2048}$.

3.2 Experiments

We trained six models to test the different methods presented in section 2. We compare the standard, and the attention-based MIL assumptions explained in 2.2 with and without contrastive pre-training of the feature extractor. We also trained two more attention-based MIL models using clinical data. Additionally, we analyzed the relative importance given to each patch by the attention model when using F_{Im} versus F_{CLR}. The results are shown in Table 2 and Figure 3 respectively.

We used a learning rate of 10^{-5} and L2 regularization (5×10^{-3}) with stochastic gradient descent as an optimizer. The specific values for learning rate, regularization, dropout probabilities, and the number of hidden layers for the clinical data input head and the classification MLPs were all selected using bayesian hyperparameter optimization. Cross entropy was used as a loss function when training the classifier. For testing, we chose the model weights with the best F1-score found in validation. Ten different random splits were used with a proportion of 80%, 10%, 10% for training, validation, and testing to obtain mean and variance.

4 Discussion

Table 2 shows that the standard MIL assumption did not yield better than random results in most computed metrics. The attention-based classifiers trained

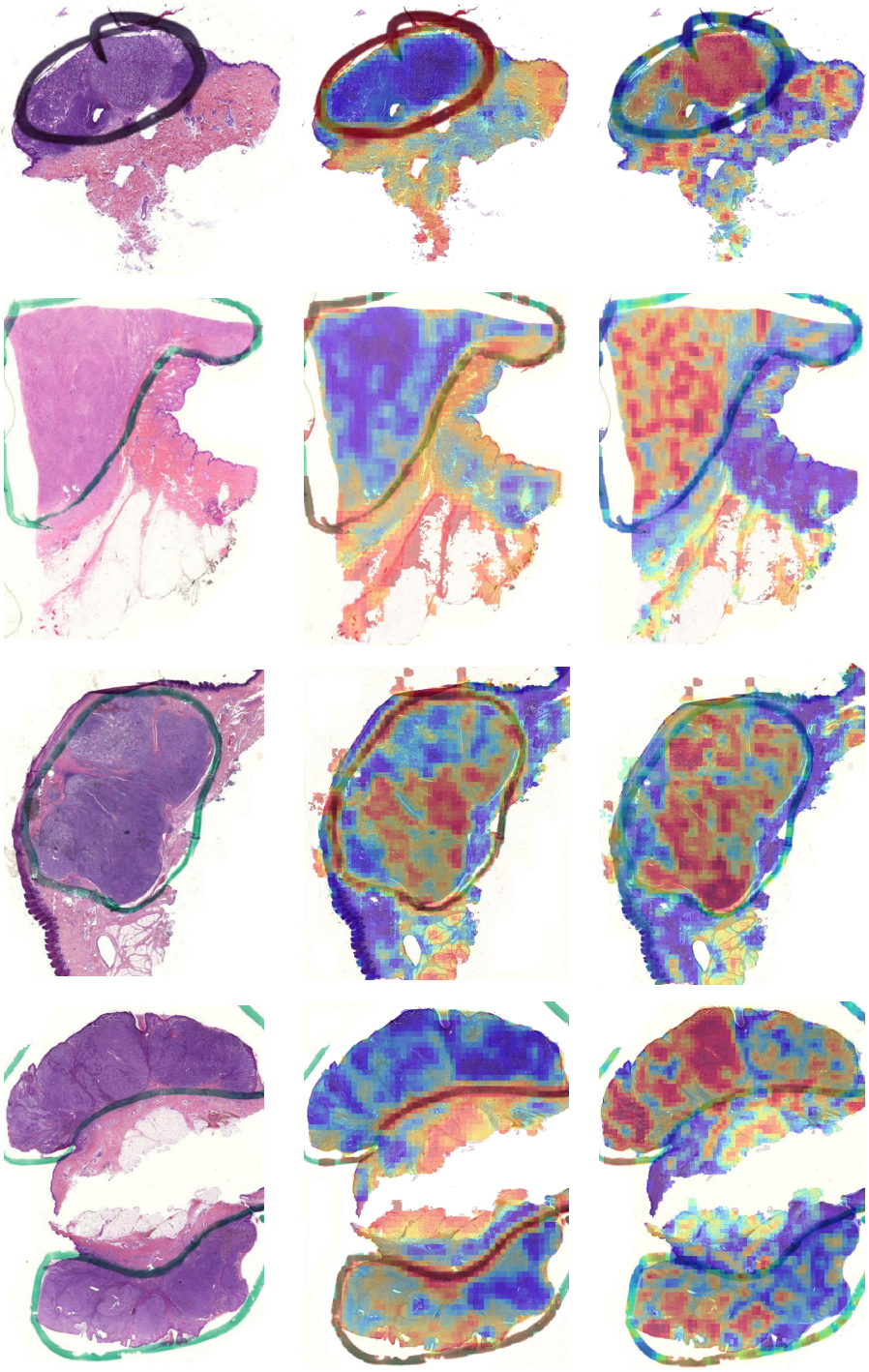

Fig. 3. Patch importance map. Original image (left), patch importance of the classifier trained on F_{Im} (center), and a second classifier trained on F_{CLR} (right).

Table 2. Quantitative comparison between the proposed models and results extracted from [3] (using a different dataset). Mean and standard deviation computed over 10 different splits.

Pre-training	F1-score	AUC	Bal. acc	Sensitivity	Specificity	Precision
Sta. MIL F_{Im}	0.51 ±0.10	0.53 ±0.13	0.50 ±0.08	**0.71 ±0.26**	0.29 ±0.28	0.51 ±0.22
Sta. MIL F_{CLR}	0.47 ±0.11	0.53 ±0.12	0.54 ±0.10	0.49 ±0.17	0.59 ±0.18	0.50 ±0.14
Att. MIL F_{Im}	0.52 ±0.14	0.55 ±0.12	0.55 ±0.13	0.60 ±0.22	0.5 ±0.19	0.51 ±0.10
Att. MIL F_{CLR}	0.57 ±0.18	0.57 ±0.23	**0.60 ±0.14**	0.69 ±0.26	0.513 ±0.25	0.53 ±0.18
Att. MIL F_{Im} & clinic	0.5 ±0.14	0.56 ±0.17	0.55 ±0.13	0.67 ±0.21	0.65 ±0.27	0.50 ±0.06
Att. MIL F_{CLR} & clinic	**0.58 ±0.15**	0.58 ±0.17	0.58 ±0.13	0.65 ±0.25	0.52 ±0.28	**0.56 ±0.15**
Brinker et al.	Not reported	**0.62 ±0.02**	0.56 ±0.02	0.48 ±0.14	**0.65 ±0.11**	0.44 ±0.01

on F_{CLR} Original WSI performed better than their F_{Im} counterpart. Finally, the addition of metadata also improved the results. Even though most reported metrics are slightly above random choice, each proposed method improved the classification capabilities of our model.

F_{CLR} also provided the classification module the capability of distinguishing between tumour and adjacent normal tissue. It also prevented the model from focalizing on artifacts of the dataset, such as pen marks drawn on the slides by the clinical practitioners. Comparing the center and right columns of Figure 3, it is clear that the attention scores on the right column are more focused on tumoural areas, while the attention scores on the center column tend to spread the patch importance across the slide or on image artifacts. This finding shows that pre-training a feature extractor with SimCLR has the potential to be used for meaningful feature extractions even in the presence of artifacts without needing manual annotation by trained dermatologists.

5 Conclusions

Our study has shown that CNN-based image classification of primary tumours to detect lymph node positivity is possible under a MIL framework. Moreover, the use of F_{SimCLR} over F_{Im}, self-attention, and clinical information allowed the model to improve all the considered metrics. The classifier trained on F_{SimCLR} has shown a qualitative improvement in the model's capability to focus on clinically relevant areas and avoid image artifacts. However, the results are still not good enough to justify using deep learning systems as a selection method for SLNBs. Further studies with more WSIs are needed to validate and increase the efficacy of the presented methods.

References

1. Akdeniz, M., Hahnel, E., Ulrich, C., Blume-Peytavi, U., Kottner, J.: Prevalence and associated factors of skin cancer in aged nursing home residents: A multicenter prevalence study. PloS one **14**(4), e0215379 (2019)
2. Barata, C., Santiago, C.: Improving the explainability of skin cancer diagnosis using cbir. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 550–559. Springer (2021)
3. Brinker, T.J., Kiehl, L., Schmitt, M., et al.: Deep learning approach to predict sentinel lymph node status directly from routine histology of primary melanoma tumours. European Journal of Cancer **154**, 227–234 (2021)
4. Chen, T., Kornblith, S., et al.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
5. Combalia, M., Codella, N., Rotemberg, V., Carrera, C., Dusza, S., Gutman, D., Helba, B., Kittler, H., Kurtansky, N.R., Liopyris, K., et al.: Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: the 2019 international skin imaging collaboration grand challenge. The Lancet Digital Health **4**(5), e330–e339 (2022)
6. Gareau, D.S., Browning, J., Da Rosa, J.C., Suarez-Farinas, M., Lish, S., Zong, A.M., Firester, B., Vrattos, C., Renert-Yuval, Y., Gamboa, M., et al.: Deep learning-level melanoma detection by interpretable machine learning and imaging biomarker cues. Journal of Biomedical Optics **25**(11), 112906 (2020)
7. Gershenwald, J.E., et al.: Melanoma staging: evidence-based changes in the american joint committee on cancer eighth edition cancer staging manual. CA: a cancer journal for clinicians **67**(6), 472–492 (2017)
8. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press (2016)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
10. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
11. Jaiswal, A., Babu, A.R., Zadeh, M.Z., Banerjee, D., Makedon, F.: A survey on contrastive self-supervised learning. Technologies **9**(1), 2 (2020)
12. Li, X., Desrosiers, C., Liu, X.: Symmetric contrastive loss for out-of-distribution skin lesion detection. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2022)
13. Lu, M.Y., Williamson, D.F., et al.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021)
14. Otsu, N.: A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics **9**(1), 62–66 (1979)
15. Suganyadevi, S., Seethalakshmi, V., Balasamy, K.: A review on deep learning in medical image analysis. International Journal of Multimedia Information Retrieval **11**(1), 19–38 (2022)
16. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **71**(3), 209–249 (2021)

17. Verdelho, M.R., Barata, C.: On the impact of self-supervised learning in skin cancer diagnosis. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2022)
18. Vij, R., Arora, S.: Computer vision with deep learning techniques for neurodegenerative diseases analysis using neuroimaging: A survey. In: International Conference on Innovative Computing and Communications. pp. 179–189. Springer (2022)
19. Wang, X., Chen, H., Gan, C., Lin, H., Dou, Q., Tsougenis, E., Huang, Q., Cai, M., Heng, P.A.: Weakly supervised deep learning for whole slide lung cancer image analysis. IEEE transactions on cybernetics **50**(9), 3950–3962 (2019)
20. Zhang, Z., Chen, P., McGough, M., Xing, F., Wang, C., Bui, M., Xie, Y., Sapkota, M., Cui, L., Dhillon, J., et al.: Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. Nature Machine Intelligence **1**(5), 236–245 (2019)