Does context matter in digital pathology?

Paulina Tomaszewska[1], Mateusz Sperkowski[1] and Przemysław Biecek[1,2]

[1] Faculty of Mathematics and Information Science, Warsaw University of Technology, Koszykowa 75, 00-662 Warsaw, Poland
[2] Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland

Abstract

The development of Artificial Intelligence for healthcare is of great importance. Models can sometimes achieve even superior performance to human experts, however, they can reason based on spurious features. This is not acceptable to the experts as it is expected that the models catch the valid patterns in the data following domain expertise. In the work, we analyse whether Deep Learning (DL) models for vision follow the histopathologists' practice so that when diagnosing a part of a lesion, they take into account also the surrounding tissues which serve as context. It turns out that the performance of DL models significantly decreases when the amount of contextual information is limited, therefore contextual information is valuable at prediction time. Moreover, we show that the models sometimes behave in an unstable way as for some images, they change the predictions many times depending on the size of the context. It may suggest that partial contextual information can be misleading.

Keywords

digital pathology, context, Deep Learning, Computer Vision

1. Introduction

Deep Learning (DL) models are perceived as black-boxes. They sometimes make decisions based on different reasons than humans do. DL models tend to take shortcuts and follow spurious correlations [1]. The popular example is the case where the model misclassified a husky as a wolf because there was snow in the background, which was a rule in the case of images with wolves within the dataset [2]. Although following such a rule may lead to high classification performance (if the dataset is biased), we expect the model to distinguish between wolves and husky dogs based on the animal features. Motivated by the fact that there is sometimes a mismatch between the way DL models and humans reason, we decided to investigate whether the DL models for vision follow the same good practices when diagnosing lesions based on histopathological data as expert histopathologists. The histopathologists when diagnosing a particular region of a lesion, take into account also the surrounding tissue [3]. We investigate whether the classification performance of DL models for vision will be higher when they have access to information about neighbouring tissues than in the case where no contextual information is given. We conduct a quantitative analysis on how the amount of contextual information within the input to the models impacts the final performance. Our contribution is as follows:

- we verify whether DL models for vision behave in a similar way to histopathologists who benefit from contextual information when diagnosing lesions
- we measure quantitatively the impact of the amount of contextual information provided to different DL models for vision on their classification performance
- we investigate whether it happens that the models behave in a non-stable way by changing predictions for the images many times given different amounts of contextual information

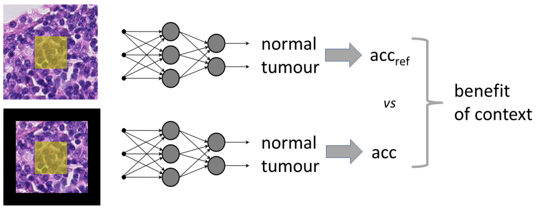

Figure 1: The scheme of the proposed study. The yellow squares in the center of histopathological images depict the regions that the annotations are based on (the squares are shown only for visualization purposes and are not present in dataset images). The black border is applied to remove some parts of contextual information.

The code to replicate our results is available at `https://github.com/ptomaszewska/PCam_context`.

MRC 2023 – The Fourteenth International Workshop Modelling and Representing Context. Held at ECAI 2023, 30.09.-5.10.2023, Kraków, Poland.

paulina.tomaszewska3.dokt@pw.edu.pl (P. Tomaszewska); mateusz.sperkowski.stud@pw.edu.pl (M. Sperkowski); przemyslaw.biecek@pw.edu.pl (P. Biecek)
0000-0001-6767-1018 (P. Tomaszewska); 0000-0001-8423-1823 (P. Biecek)

2. Motivation

The histopathological data is saved in the form of Whole Slide Images (WSIs) where the whole lesion is under huge resolution. In the popular Camelyon16 Breast Cancer dataset [4] the images have an average resolution of 94,747x188,764 (avg. size of 1.97GB). Such huge images are difficult to load into memory and process, therefore they are most commonly split into smaller-sized patches. The advantage of WSIs is that based on them both local (cell-level) and global (tissue-level) analysis can be performed. In the work, we use the variant of the Camelyon16 dataset, PatchCamelyon (PCam) [5], where the initial WSIs are cut into patches of size 96x96. Each such patch has a label - normal tissue or tumour lesion and the model's task is a binary classification. However, in such a dataset, the global context is not preserved as the relationship between neighbouring patches is lost. Nevertheless, the dataset contains local context within images since the label is assigned to the whole patch only based on the central 32x32 region of the patch. Therefore, the surrounding box can be thought of as contextual information. The question is whether contextual information is useful when making predictions.

3. Method

The goal of the study is to check whether the DL models for vision are sensitive to different amounts of contextual information within input histopathological images. First, we use the test set of the PCam dataset as input for inference to the already trained models on the original PCam dataset. The resulting performance metrics serve as a reference point when the full available context is provided to the model (acc_{ref}, $precision_{ref}$, $recall_{ref}$, AUC_{ref}).
In the following experiments, we restrict the amount of contextual information in the image data. Let us define, the size of context (s) as a width in pixels of the area around the central 32x32 square. The maximum size of the context is $(96 - 32)/2 = 32$ where 96 is the length of the original image's side. The bigger the context size, the more information about the neighbouring tissue is provided to the model. To evaluate how the size of the context impacts the model prediction, we remove external layers of pixels of the context area and the image is padded to the original size of 96x96 with black pixels (we call it a border for brevity). We decided to use the black colour as it is often used as a baseline colour in explainable AI (XAI) methods i.e. Integrated Gradients [6]. We applied the border to the images to obscure some part of the context instead of cutting off the pixels and changing the image resolution to avoid a situation where it would be difficult to disentangle the source of the performance change - increased resolution vs. limited context size. The images padded with a black border are the input to the DL models. We analyse the difference between the metrics obtained on padded images and the reference, original ones. This will give us information on how much contextual information is beneficial when making predictions.

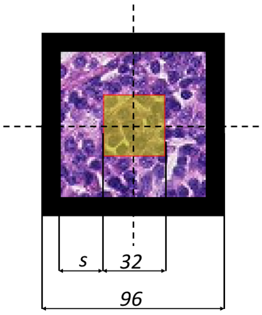

Figure 2: The histopathological image padded with black border with dimensions specified. The dimension s denotes context size.

4. Experiments

4.1. Deep Learning models

We apply the method described in Section 3 on two classes of DL models - convolutional (*ResNet18* and *DenseNet121*) and transformer-based (*Swin* [7] and *ViT* [8]). The convolutional models trained on histopathological data (without any image standarization) were taken from [9]. In the case of the transformer-based models, we took the models pretrained on Imagenet. The *Swin* model was pretrained in a supervised manner, whereas in the case of *ViT*, we used the models pretrained in different schemes: supervised (called *supViT* for brevity) and unsupervised (contrastive - *MoCo* [10] and autoencoder-based - *MAE* [11]). We applied end-to-end finetuning using the whole PCam training set. As the pretrained transformer-based models operate on an input size of 224x224 and the images within PCam dataset are of size 96x96, we applied resizing (as it is done in [12]). We did not apply any standarization to keep the same preprocessing procedure as in the convolutional models. The hyperparameters such as base learning rate (lr_{base}) and weight decay (wd) were the same as the ones used by the authors of [12] (details in Table 1). Compared to the paper, we decreased the batch size from 128 to 64 and shortened the training procedure. The base learning rate was linearly scaled using the formula $lr = lr_{base} * batch\_size/256$ [13]. The models were finetuned for 5 epochs with Adam optimizer and with a linear learning rate scheduler increasing to the final lr value. We took the model from the epoch that resulted in the biggest accuracy on the validation set. The fact that a small number of epochs was enough to achieve satisfactory accuracy was due to the big size of the training set (262,144 samples).

Table 1
Hyperparameters used for finetuning of transformer-based models on PCam dataset.

model	lr_{base}	*wd*
Swin	0.0001	0.001
supViT	0.001	0.01
MAE	0.0001	0.0001
MoCo	0.0001	0.0001

5. Results

In the PCam dataset, there are duplicates of images due to the procedure used to create it. In all experiments, we use the test set after the removal of the duplicates. First, we perform the inference using a test set with original images (with full context size) using the analysed models. As a result, the reference values of performance metrics are obtained (Table 2).

Table 2
Deep Learning models' performance when full available contextual information is given (so-called reference performance).

model	acc_{ref}	$precision_{ref}$	$recall_{ref}$	AUC_{ref}
ResNet18	0.8786	0.9396	0.7754	0.9477
DenseNet121	0.8941	0.9498	0.8029	0.9904
Swin	0.9172	0.9680	0.8405	0.9716
supViT	0.9160	0.9514	0.8537	0.9734
MAE	0.9173	0.9513	0.8568	0.9746
MoCo	0.9143	0.9617	0.8397	0.9708

The analysed models have similar reference performance. However, it is observed that transformer-based models perform slightly better than convolutional models. Note that the convolutional models have smaller capacity as they have much less tunable parameters in millions (*ResNet18* - 11.7, *DenseNet121* - 8) than transformer-based models (*Swin* - 86.7, *ViT* - 85.8), which can be a source of the difference.

5.1. Image dimensions mapping

As already mentioned transformer-based models operate on the resized PCam images (224x224) whereas the convolutional models - on the original size of the images (96x96). Therefore, the maximum size of the context and the central square is increased. This mismatch is important when analysing the results on a common axis. Note that the mapping from the size of 96x96 to 224x224 is of factor 2.33 which makes the alignment difficult when trying to map the values of context size in the two scenarios. As a solution, we decrease the context size in the resized images by 7 pixels as opposed to the scenario with images of original size where it is done pixel by pixel. Cutting out the context by every 7th pixel in the resized image translates to cutting every 3rd pixel in the original image. As a consequence in the plots in the following sections, there will be more points on the curves referring to the convolutional models than the transformer-based ones. Note that the context size of 32 within the original images maps to the value of 74.67 within the resized images, therefore, between the 74th and 75th pixel (which we additionally analyse despite the policy to cut off the context by every 7th pixel). To aggregate the results from the two context sizes, we apply a conservative approach. We take the score that is the furthest from the label meaning that (1) if the correct class is 1, we take the lowest probability from the two pixels (74th and 75th), (2) if the label is 0, we take the biggest probability.

5.2. Drop of performance with the decrease of context size

Having reference values of performance metrics, we limit the context size within the input images by applying a black border to hide parts of the contextual information. The performance metrics for different models are shown in Figure 3.

It is observed that the metric that is the most affected by limiting the amount of contextual information available at the prediction time is *recall*. This metric is especially important in healthcare where false negatives are of the greatest concern. The biggest drop in *recall* (0.84) occurs in the case of *MoCo* model which starts to predict mostly one class - normal. At the same time, in the case of this model, the *precision* curve slightly increases with the decrease of context size (up to 0.016 when $context\_size = 0$) even though the reference value was already very high (0.9617). The two models that experience the smallest drop of *recall* are *Swin* (0.41) and *DenseNet121* (0.40).

Note that *precision* seems to be the least impacted metric by the limitation of contextual information. The most considerable decrease is observed in the case of *DenseNet121* (0.14) and *MAE* (0.11), however, in the first case, the significant drop is observed even when the amount of context is only slightly limited.

In the case of *accuracy* and *AUC* plots, all the models behave similarly except *MoCo* which has the biggest drop of about 0.41. Interestingly, the *accuracy*, *recall* and *precision* curves of *ResNet18* experience significant fluctuation over different context sizes. In the case of *accuracy* and *recall*, local minimum is observed for a context size of about 24, whereas the local maximum at 8.

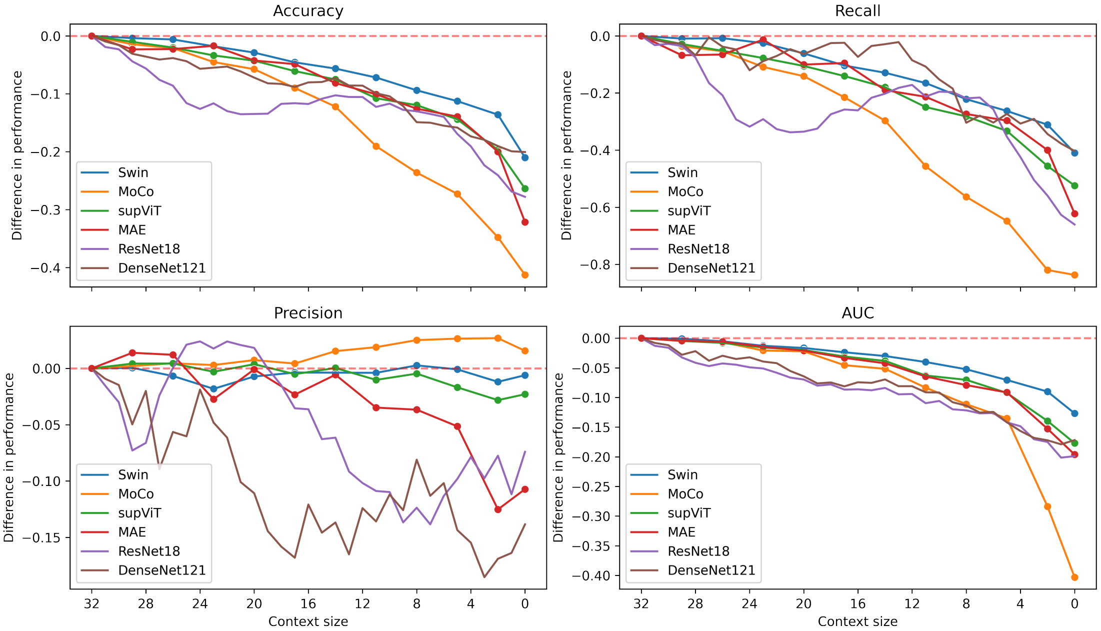

Figure 3: The performance gap when the context size is limited. The performance metrics of Deep Learning models decreased by the respective reference values (when full context is available) under different context sizes are shown. Note that the values on the x-axis are in decreasing order which makes an interpretation of the experiments easier. The y-axis is not shared within the subplots so that the variations of results for different models are more visible. The markers on the curves corresponding to transformer-based models highlight a smaller number of data points than in the case of convolutional models.

Note that by the fact that fewer data points are depicted in the case of transformer-based models than in the convolutional models, the smoothness of the curves cannot be compared between the two families of models, unlike the general trends. Overall, it seems that the analysed convolutional models are more sensitive to the lack of contextual information than the transformer-based models (except *MoCo*). However, the differences do not seem significant taking into account the gap in the models' capacity.

5.3. Consistency of predictions over different context sizes

Despite the changes in models' performance, the predictions for some images are the same over different context sizes. It turns out that the model that generates the consistent predictions most often is *Swin* and at the second place - *supViT* (see Table 3). The rest of the models are far behind when it comes to the percentages of images with agreeing predictions. The most sensitive to the decrease of the context size is *DenseNet121*.

Table 3
Percentage of images that get the same prediction regardless of the context size.

ResNet18	*DenseNet121*	*Swin*	*supViT*	*MAE*	*MoCo*
54.66	52.39	72.41	69.06	60.76	58.20

5.4. Misleading pieces of context

The decrease in performance results means that the models change their predictions for some images depending on the amount of contextual information. We distinguish two types of images (1) the ones that undergo the change of prediction given a particular model only once (2) the ones that experience the change of prediction more than once (referred to as *'swinging images'*). We present how the probability of class tumour changes depending on the context size for the sample images from the aforementioned two types in the case of *DenseNet121* (Figure 4). It is seen that the images that undergo frequent changes of the predictions are the ones that get the probabilities close to the threshold of 0.5. However, note that there is one sample (depicted in blue colour) that the model was initially confident about (probability above 0.8) but with the decrease of the context size, the probability went down and started to oscillate around the threshold. It can be seen that in the case of images that experience the change of prediction only once, it happens for different context sizes.

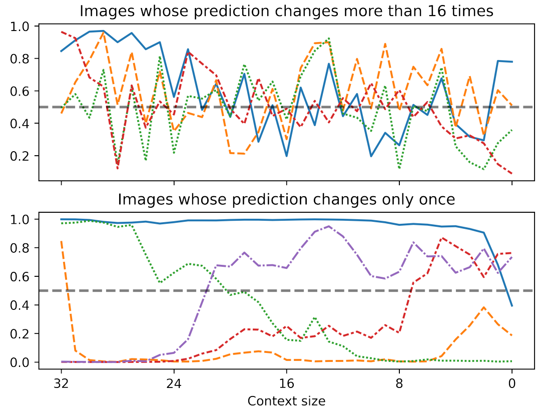

Figure 4: The misleading nature of context depending on its size (on the example of *DenseNet121*). The probability of the class tumour is shown in two cases: all images that experienced the change in the prediction in more than half of all context sizes (top), and a sample of images that changed the class only once (bottom).

The analysis of when (at what context size) the model changes predictions was performed. We analyse the number of images undergoing the particular direction of a change for the first time at the given context size. In the analysis, the *'swinging images'* are not included. We show results only for *DenseNet121* as for illustration (Figure 5). Interestingly, it can be seen that the changes of predictions occur mostly for the extreme values of context size - either the small or the big ones. This observation holds for other models except *ResNet18* where the shift TN->FP happens mostly for moderate values of context size, however, the scale of the shift is small as for the given moderate context size there are a maximum of 11 images that undergo a particular change of prediction.

5.4.1. Possible interpretation

The shift FP->TN could suggest that the model has seen some tumour cells in the context area as it predicted the class tumour but by limiting the amount of contextual information, we cut off the pixels containing these tumour cells and the model outputs the correct class 'normal'. Such behaviour of the model could be understandable as the model was not explicitly said to base its opinion only on the central square.

The other shifts are more difficult to find a reasonable explanation of. For example, the shift FN->TP could mean that initially, the model focused too much on contextual information where there was only normal tissue. When this distraction in the form of contextual information was taken away, the model paid only attention to the key central part and spotted the tumour cells and as a consequence output the correct class. However, such a behaviour of the model is not desirable.

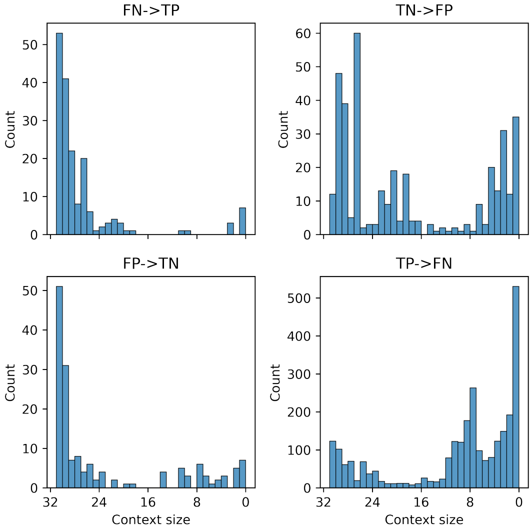

Figure 5: Number of images undergoing the change of prediction for the first time given the particular context size with the distinction on the initial and consecutive model prediction (*'swinging images'* not included). Note that for better visibility, the y-axis is not shared between subplots. The bin width is equal to 1. The results are provided for the *DenseNet121*.

In the aforementioned possible interpretations, we focused on the ones that do not require domain knowledge. However, the tissue structure and some spatial relationships that were partially covered by the black borders may potentially also play a role in the changes of model predictions.

5.4.2. Summarized model behaviour

The summarized results from the histograms (without the distinction on the type of prediction shift) corresponding to different models are shown in Figure 6. It turns out that indeed (as shown in Figure 5) in the case of *DenseNet121*, the most changes occur for extreme context sizes but it is even more visible in the case of *ResNet18*. In *supViT* and *MAE*, the biggest boost of the number of changes occurs when the context size is significantly reduced. The most changes in total are observed when *MoCo* is used.

Lastly, it was analysed how many images in total experience a particular number of prediction changes given a particular model (Figure 7). It turns out that the models that change predictions more than 24 times per im-

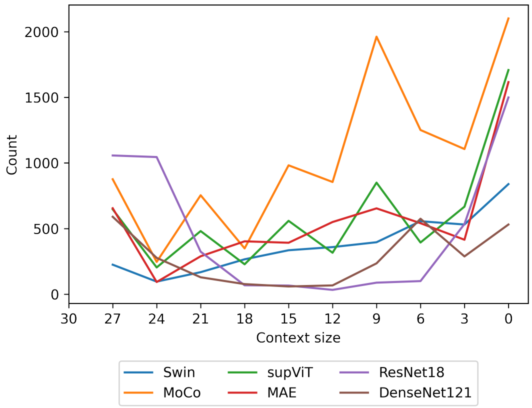

Figure 6: Number of images undergoing the change of prediction for the first time given the particular context size (*'swinging images'* not included). To account for the fact that in the case of transformer-based models, we cut the context size by every 3 pixels (after mapping the image size back to the original one) not by 1 pixel as in the case of convolutional models (see Section *Image dimensions mapping*), we summed the number of images undergoing the change from three consecutive context sizes. By the application of the aforementioned 'normalization', the curves of convolutional and transformer-based models are on the same scale and have the same number of data points).

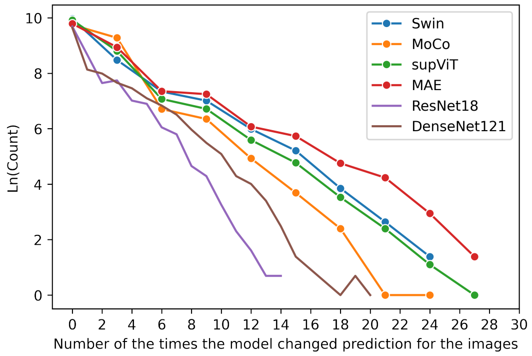

Figure 7: The logarithm of the number of images that experience the particular number of prediction changes given a particular model. To account for the fact that the maximum possible number of prediction changes in the case of transformer-based models was 10 and in the case of convolutional models - 31, the values on x-axis are multiplied by the factor of 3 (for transformer-based models) to put them on the same scale as convolutional models allowing a fair comparison.

age (out of 30 analysed possibilities, after rescaling explained in the Figure's caption) are *MAE* and *supViT*. However, these are very rare - in the case of *MAE*, there are only four such images, and in the case of *supViT*- only one. Note that the area under the curves cannot be compared between the models from the convolutional and transformer families as in the latter there are fewer data points. However, when comparing *DenseNet121* and *ResNet18*, it is visible that for the *'most swinging'* images, the changes of predictions are more frequent in the case of *DenseNet121* than *ResNet18*. The four *'most swinging'* images given *DenseNet121* are shown in Figure 4. The most similar models in behaviour are *supViT* and *Swin* where the relationship between the logarithm of the number of images undergoing the particular number of changes and the number of changes is almost linear.

5.4.3. Models' agreement

It is analysed whether the same images are confusing to the models when the context size is limited. We investigate how many *'swinging images'* are in common for any pair of models regardless of context size when the changes of predictions occur. We analyse transformer-based models and convolutional ones separately. It turns out that the biggest agreement between the transformer-based models is in the case of *Swin* and *MAE* (1176 cases) whereas the smallest agreement is between *supViT* and *MoCo* (502 cases). In the case of *ResNet18* and *DenseNet121*, there are 4822 *'swinging images'* in common, therefore, they seem much more alike than transformer-based models even though three out of four analysed transformer-based models have the same architecture (ViT) but differ in pretraining scheme.

6. Conclusions

In the work, we investigate whether the Deep Learning models for vision are sensitive to contextual information when making predictions on histopathological data. It turns out that when the context size is limited, the models achieve worse performance than in the case when full context is available which means that context is important at prediction time. It is observed that depending on the amount of contextual information, the model can output different predictions for a given image. We evaluate the behaviour of models that have similar reference performance metrics (when full access to a context is provided) in the case when the size of contextual information is decreased. It turns out that the model that is the most sensitive to the limitation of context size is *MoCo*. It may possibly be attributed to the fact that the model was pretrained in a contrastive way but it requires further investigation. We observe that there are images of two types - the ones that undergo one change of prediction and the *'swinging images'*. For the latter, it would be interesting to consult the images with a histopathologist to verify whether indeed in these cases, the context may be misleading. Moreover, in the future, the possible interpretations of the obtained results could be complemented by the analysis with the use of heatmap-based XAI techniques.

Acknowledgments

The project was funded by POB Research Centre Cybersecurity and Data Science of Warsaw University of Technology within the Excellence Initiative Program - Research University (ID-UB). This research was carried out with the support of the Laboratory of Bioinformatics and Computational Genomics and the High Performance Computing Center of the Faculty of Mathematics and Information Science Warsaw University of Technology.

References

[1] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, F. A. Wichmann, Shortcut learning in deep neural networks, Nature Machine Intelligence 2 (2020) 665–673. doi:10.1038/s42256-020-00257-z.

[2] M. Ribeiro, S. Singh, C. Guestrin, “Why Should I Trust You?”: Explaining the Predictions of Any Classifier, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, Association for Computational Linguistics, San Diego, California, 2016, pp. 97–101. URL: https://aclanthology.org/N16-3020. doi:10.18653/v1/N16-3020.

[3] K. Zormpas-Petridis, H. Failmezger, S. E. A. Raza, I. Roxanis, Y. Jamin, Y. Yuan, Superpixel-Based Conditional Random Fields (SuperCRF): Incorporating Global and Local Context for Enhanced Deep Learning in Melanoma Histopathology, Frontiers in Oncology 9 (2019). doi:10.3389/fonc.2019.01045.

[4] B. Ehteshami Bejnordi, M. Veta, J. P, al, F. Beca, S. Albarqouni, R. Cetin-Atalay, T. Qaiser, I. Serrano Gracia, M. Shaban, A. Kalinovsky, H. Matsuda, S. Seno, K. Kartasalo, D. Racoceanu, Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer, JAMA 318 (2017) 2199–2210. doi:10.1001/jama.2017.14585.

[5] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, M. Welling, Rotation Equivariant CNNs for Digital Pathology, in: A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, G. Fichtinger (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, Springer International Publishing, 2018, pp. 210–218.

[6] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML’17, JMLR, 2017, p. 3319–3328.

[7] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 9992–10002.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, Internation Conference on Representation Learning (ICLR) (2021).

[9] J. Pocock, S. Graham, Q. D. Vu, M. Jahanifar, S. Deshpande, G. Hadjigeorghiou, A. Shephard, R. M. S. Bashir, M. Bilal, W. Lu, D. Epstein, F. Minhas, N. M. Rajpoot, S. E. A. Raza, TIAToolbox as an end-to-end library for advanced tissue image analytics, Communications Medicine vol. 2 (2022) 120. doi:10.1038/s43856-022-00186-5.

[10] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum Contrast for Unsupervised Visual Representation Learning, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, 2020, pp. 9726–9735. doi:10.1109/CVPR42600.2020.00975.

[11] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, R. Girshick, Masked Autoencoders Are Scalable Vision Learners, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, 2022, pp. 15979–15988. doi:10.1109/CVPR52688.2022.01553.

[12] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, S.-N. Lim, Visual Prompt Tuning, in: European Conference on Computer Vision (ECCV), 2022.

[13] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, K. He, Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour (2018). arXiv:1706.02677.