Evaluation and Comparison of CNN Visual Explanations for Histopathology

Mara Graziani[1,3,*], **Thomas Lompech**[2,*], **Henning Müller**[1,3], **Vincent Andrearczyk**[1]
[1]University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland
[2]INP-ENSEEIHT, 31000 Tolouse, France
[3]University of Geneva, Geneva, Switzerland
[*]Equal contribution

Abstract

Visualization methods for Convolutional Neural Networks (CNNs) are spreading within the medical community to obtain explainable AI (XAI). The sole qualitative assessment of the explanations is subject to a risk of confirmation bias. This paper proposes a methodology for the quantitative evaluation of common visualization approaches for histopathology images, i.e. Class Activation Mapping and Local-Interpretable Model-Agnostic Explanations. In our evaluation, we propose to assess four main points, namely the alignment with clinical factors, the agreement between XAI methods, the consistency and repeatability of the explanations. To do so, we compare the intersection over union of multiple visualizations of the CNN attention with the semantic annotation of functionally different nuclei types. The experimental results do not show stronger attributions to the multiple nuclei types than those of a randomly initialized CNN. The visualizations hardly agree on salient areas and LIME outputs have particularly unstable repeatability and consistency. The qualitative evaluation alone is thus not sufficient to establish the appropriateness and reliability of the visualization tools. The code is available on GitHub at `bit.ly/2K48HKz`.

In many medical imaging tasks, such as segmentation and classification in magnetic resonance, computed tomography or ultrasound images, the input comprises an image with a clear region of interest (e.g. organ or tumor) that often comes with ground-truth segmentation. In histopathology, Whole Slide Images (WSIs) can reach gigapixel sizes, with sometimes only infinitesimal regions being decisive for the task. Isolated tumor cells, for example, can determine crucial decisions despite being single cells or small clusters of tumor cells that occupy less than 0.008% of the entire image.

Understanding the decision-making process of Convolutional Neural Networks (CNNs) is a key point in medical imaging, to ensure that clinically correct decisions are taken. Among the explainability (XAI) methods proposed in the literature, the post-training attribution to either high-level concepts (Graziani, Andrearczyk, and Müller 2018; Graziani et al. 2020; Kim et al. 2018) or input features was proposed for medical applications (Palatnik de Sousa, Maria Bernardes Rebuzzi Vellasco, and Costa da Silva 2019; Ribeiro, Singh, and Guestrin 2016). Feature attribution, in particular, highlights the most influential set of features in the input space by generating saliency maps, also called heatmaps (Selvaraju et al. 2017; Palatnik de Sousa, Maria Bernardes Rebuzzi Vellasco, and Costa da Silva 2019; Chattopadhay et al. 2018; Ribeiro, Singh, and Guestrin 2016; Zhou et al. 2016). One of the risks of accepting the plausibility of the heatmaps only by visual assessment is that of incurring in the so-called confirmation bias. As the research in cognitive psychology explains, we tend to attribute greater confidence to a hypothesis, even if false, when explanations are generated for it (Lombrozo 2006). For this reason, the reliability and trustworthiness of visual explanations should be thoroughly addressed before their incorporation into healthcare pipelines, clarifying the advantages, limitations and similarities of the methods. As Tokenaboni argues in (Tonekaboni et al. 2019), the specific needs of clinical practice require the evaluation of the appropriateness of the explanations, their alignment with clinical factors, their potential of being translated into action and, finally, their consistency over parameter shifts. Remarkable evaluations of the consistency of saliency maps proposed adding constant shifts into the data (Kindermans et al. 2019), comparing visualizations after cascading randomizations of the network weights (Adebayo et al. 2018) and quantifying the similarity of explanations under multiple conditions (Arun et al. 2020). The instability of XAI visualization methods applied to natural images and chest X-rays emerged from these studies. If the lesion contours are available, the appropriateness of the explanations can be evaluated by localization metrics (Arun et al. 2020). Evaluating visualization methods on the basis of their localization performance as in (Arun et al. 2020), however, may easily fail in the context of histopathology images. WSIs do not have a clear central subject on the foreground but rather a structural disposition of many instances (e.g. connective, adipose, or epithelium cells) at several scales, as illustrated in Fig. 1.

In this work, we propose quantitative metrics that can specifically evaluate visual explanations in the context of histopathology images (at 40X magnification). To establish whether the visualizations are appropriate for the domain, we evaluate the Intersection over Union (IoU) between the heatmaps and functionally different nuclei types, i.e. neo-

plastic, inflammatory, epithelial and connective nuclei. We then compare the Structural SIMilarity index (SSIM) between multiple XAI methods in order to help users with choosing the best visualization method and to reduce the number of look-alike visualizations. The consistency of the explanations is evaluated against small shifts in the hyper-parameters of the explanation techniques, and against the cascading randomization of the model parameters (i.e. CNN weights). We finally assess, where needed, the repeatability of the explanations for multiple initialization seeds. Our analysis evaluates the commonly used activation maps and linear surrogate models XAI methods, namely Class Activation Mapping (CAM), its gradient-based evolutions Grad-CAM and Grad-CAM++, and Local-Interpretable Model Agnostic Explanations (LIME). Results on the Camelyon and PanNuke data collections show that the XAI visualizations in this paper do not explain the CNN decisions in terms of the attention paid to tumorous nuclei. The visualizations, moreover, disagree on the salient input regions, and appear inconsistent for small shifts in the XAI hyper-parameters.

Methods

Breast Tissue Classification

Datasets We use a combination of three publicly available datasets, namely Camelyon 16, Camelyon 17 (Litjens et al. 2018) and the breast subset of the PanNuke dataset (Gamper et al. 2019)[1].

The Camelyon collection includes more than a thousand training WSIs of lymph node sections (899 for Camelyon 17 and 270 for Camelyon 16), with slide-level annotations of metastasis type (negative, macro-metastases, micro-metastases, isolated tumor cells) and a few manual segmentations of tumor regions (only for 320 WSIs). The data were collected at five medical centers with three scanner types and present high heterogeneity and staining variability (Khan et al. 2020).

The PanNuke dataset is a collection of 481 WSIs from 19 different tissue types with semi-automatic instance segmentations of five different nuclei types, namely neoplastic, inflammatory, connective, epithelial and dead nuclei. No dead nuclei were segmented for the breast data used in our work, as shown by the nuclei type statistics in (Gamper et al. 2020). The benefit of combining images from several datasets is twofold. It improves the model generalization to unseen data and, at the same time, it provides semantic nuclei segmentations to evaluate the overlap of the visual explanations with various nuclei types.

Image patches (of 224×224 pixels, i.e. the network input size) are extracted from all WSIs at 40X magnification, the highest level of magnification, to show the qualitative features of the nuclei that are prognostic of cancer (Rakha et al. 2008). Since the inputs from PanNuke are already released in image patches of size 256×256 pixels, and since they are under-represented in number with respect to the Camelyon dataset, we oversample the original image patches to a five times larger dataset. The oversampling is obtained by cropping the images at several locations (center, upper left, upper right, bottom left and bottom right corners). The training, validation and testing splits are summarized in Table 1. For the training split, the PanNuke patches were extracted from the first two original splits provided for benchmarking the data. The third split was used as testing data. Thus, the training set comprises 152,296 images (of which 123,820 do not contain tumor, i.e. negative and 28,476 contain tumor, i.e. positive). 1,820 images from two centers of the Camelyon17 data were used as validation. Finally, the test set comprises a total of 6,589 images with 2,690 negative and 3,899 positive samples. Reinhard normalization is applied to all the patches to reduce the stain variability, as suggested in (Khan et al. 2020).

Table 1: Summary of the training, validation and testing splits.

		Cam 16	Cam 17	Pannuke
Train	Negative	12,954	107,951	2,915
	Positive	6,036	17,475	4,965
Validation	Negative	-	820	-
	Positive	-	1000	-
Test	Negative	-	1,215	1,475
	Positive	-	1,499	2,400

Network Architecture and Training The CNN architecture is an Inception V3 (Szegedy et al. 2016) with ImageNet pre-trained weights that is entirely finetuned on the histopathology training images to solve the binary classification task of distinguishing positive samples against negative ones. This solution outperforms other architectures, and is therefore used for the analyses. Three fully-connected layers (2048, 512 and 256 neurons respectively) with dropout probability of 0.8 and a prediction layer are added on top of the pre-trained features. The weighted binary cross-entropy loss is used to address the strong class imbalance in the training data. L2 regularization is used with a coefficient of 0.01 on the fully-connected layers. The optimization is solved with stochastic gradient descent and standard parameters (Nesterov momentum at 0.9 and decay at 1e-6). Early stopping is performed on the validation loss to stop the training process, with 5 epochs of patience. The network obtains a test accuracy of 0.80 with an Area Under the ROC Curve (AUC) of 0.85.

Visualization Methods

Class Activation Maps Three types of activation maps are used for the analysis, namely the original CAM implementation (Zhou et al. 2016), the gradient-weighted CAM known as Grad-CAM (Selvaraju et al. 2017) and its generalized version Grad-CAM++ (Chattopadhay et al. 2018). CAM produces a localization map by visualizing the contribution of each feature map before these are spatially averaged and linearly combined to produce the network prediction. Grad-CAM generalizes CAM as it generates visual explanations by directly taking into account the cascade of

[1] https://camelyon17.grand-challenge.org/ and https://jgamper.github.io/PanNukeDataset/

gradients, thus applying to a wider variety of models and applications, including image captioning and query answering. These two methods were shown to be equivalent up to a normalization constant that is proportional to the number of pixels in the feature maps (Selvaraju et al. 2017). As a further development, Grad-CAM++ considers the gradients at the pixel level rather than those of the entire feature maps[2]. As a result Grad-CAM++ explanations partially address the shortcomings of considering the entire feature maps, like the difficulty to operate when multiple occurrences of instances of the same class occur in a single image.

LIME Interpretable surrogates are used as linear classifiers by LIME to locally simulate the decisions of the global classifier. Using LIME to explain CNNs is similar to using a sparse linear model to approximate the complex decision function of the CNN. The first step of the application of LIME to images consists of clustering pixels into superpixels (that will be used as features) using color, texture and other types of local similarities. Randomly hiding some of the superpixels generates perturbations (called samples) of the original images which can be used to compute the relevance of each superpixel. Following the indications in (Palatnik de Sousa, Maria Bernardes Rebuzzi Vellasco, and Costa da Silva 2019), we test two superpixel algorithms namely Simple Linear Iterative Clustering (SLIC) (Achanta et al. 2012) and Felzenszwalb's graph based image segmentation (FHA) (Felzenszwalb and Huttenlocher 2004).

Evaluation Methods

Visual Similarity and Alignment with Clinical Factors For the methods considered in this analysis, namely CAM, Grad-CAM, Grad-CAM++, and LIME with SLIC and FHA superpixels, we propose the qualitative evaluation of some visualizations and two quantitative analyses. The quantitative analyses measure respectively the accordance of methods (in terms of their visual similarity) and their alignment with clinical factors. In the first quantitative analysis, the SSIM is used to establish whether different XAI methods point to the same input regions to explain a given prediction. The SSIM, ranging from 0 (no structural similarity) to 1 (identical structural similarity), is computed for pairs of XAI methods to evaluate their agreement. In the second analysis, we follow the experiments in (Zhou et al. 2018) to compute the overlap (IoU) of the explanations with specific image regions, i.e. each of the four nuclei types. Particular attention is given to neoplastic nuclei as they are indicators of tumor tissue (Gamper et al. 2019; 2020).

Consistency and Repeatability The consistency of the visualizations is evaluated over variations in the method hyper-parameters. The number of superpixels and the size of the neighborhood considered by the linear surrogate (the number of samples used to solve the local linear classification task in LIME) are hyper-parameters that need to be tuned to generate meaningful LIME explanations. By using SSIM, we evaluate the similarity of the visualizations obtained for slightly increasing values of the hyper-parameters, until a visible plateau is reached (meaning that no further change happens with an additional increase in the hyper-parameter value). Based on the observations in (Madhyastha and Jain 2019), we further evaluate the impact of the inherent randomness within the LIME explanation. We compute the repeatability as the SSIM between visualizations generated for multiple initialization seeds of the surrogate model. The initialization seed controls both the random choice of the local neighborhood samples and the starting point of the optimization algorithm for the training of the local surrogate. CAM visualizations are excluded from this analysis as they do not depend on the hyper-parameter choice.

Randomization Test The randomization test verifies the dependence of the explanation on the model parameters. The output of the explainability obtained from a trained network is compared to that of a network with some randomly initialized parameters. If no clear change is present between the explanation of the trained CNN and that with randomly initialized weights, then no clear link can be established between the network weights and the explanation. This can lead to a misleading interpretation of the visualizations. With the cascading randomization test (Adebayo et al. 2018), the CNN weights are randomized in progression from the top layer to the bottom one. For each layer we evaluate the similarity between the original explanation and the one obtained with random weights up to that layer. As previous work has shown, the SSIM of explanations for ImageNet inputs does not necessarily decrease with the progressive randomization of the InceptionV3 layer (Adebayo et al. 2018). From the cascading randomization, we can only assess if the heatmaps depend on the network weights, as in the Pneumothorax examples in (Arun et al. 2020) but we cannot establish whether, after training, the heatmaps point more towards neoplastic nuclei than before. Therefore, we compute as an additional analysis the IoUs with each nuclei type for a fully randomized untrained CNN. This analysis is conducted to highlight whether the explanations become more aligned with clinically relevant factors after training.

Results

Visual Similarity of XAI Methods

We compare the visualizations obtained for test PanNuke images in Fig. 1. In this visual inspection, we compare the heatmaps of four images with different classification outcomes, namely True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN)[3]. The semantic segmentation of the nuclei is overlayed on the original images. To enable a fair comparison across the different methods and the different inputs, the heatmaps are normalized between zero and one according to the maximum

[2]The Rectified Linear Unit (ReLU) activation function that is applied to the gradients is neglected in the original formulation (Chattopadhay et al. 2018). The backpropagation of the gradients is as a result not thresholded by the activation function.

[3]More images are available in the Github repository for further validation: `https://bit.ly/2K48HKz`.

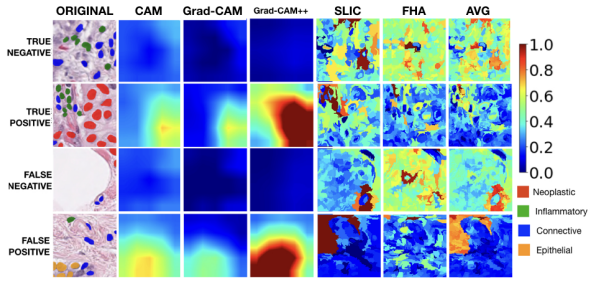
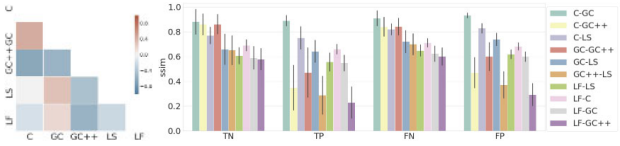


Figure 1: Qualitative comparison of CAM, Grad-CAM, Grad-CAM++ and LIME on testing images. We report LIME computed with SLIC and FHA superpixel extraction as well as their average. The nuclei segmentations are overlayed on the original images in the left column.



(a) (b)

Figure 2: a) Pearson's correlation between the SSIM of pairs of XAI methods on testing input images and the relative CNN predictions. For all cells, except C/LS, $p << 0.005$. b) Average SSIM between pairs of XAI methods for the different network outcomes, i.e. TN: True Negative, TP: True Positive, FN: False Negative, FP: False Positive. The error bars represent the standard deviation of the SSIM values.

and minimum values of the heatmaps for all testing inputs. As suggested in (Palatnik de Sousa, Maria Bernardes Rebuzzi Vellasco, and Costa da Silva 2019), the average of the LIME visualizations obtained with the two superpixel extraction algorithms is also shown for qualitative assessment (LIME AVG). To obtain LIME heatmaps, we use the maximum number of features for the explanations, corresponding to using all the superpixels in the images. We set the neighborhood size to 10,000 samples to obtain more robust visualizations, despite the high computational cost of the operations (Ribeiro, Singh, and Guestrin 2016). Heatmaps of negative predictions (both TNs and FNs) have lower values than those for TPs. The mean value of the CAM values, for example, are 1.53 for TNs, 1.83 for FNs and 4.03 for TPs.

A question that may arise is whether the heatmaps agree (i.e. high SSIM) when the network is confident about the predicted class. We evaluate the correlation between the CNN predictions and the SSIM values to see whether the similarity increases for increasing values of the prediction. The SSIM is computed for 200 randomly drawn test inputs with class stratification (100 for the positive class and 100 for the negative class). Pearson's correlation of the SSIM values is shown in Fig. 2a. Grad-CAM behaves as a generalization of CAM (Selvaraju et al. 2017), and their agreement is positively correlated with the network confidence (with Pearson's correlation coefficient $\rho = 0.44$ ($p << 0.001$). The correlations between the prediction and the SSIM of the other pairs of XAI methods in Fig. 2a are mostly negative, showing that heatmaps are more likely to disagree if the prediction is positive.

In Fig. 2b, we compare the average SSIM values for pairs of XAI methods. The 200 testing inputs are divided according to their classification outcomes, i.e. TP, TN, FP, FN. The XAI methods agree more on negative predictions than on positive ones, with an SSIM above 0.6 for all methods mostly due to consistently low activations of the heatmaps.

Alignment with Clinical Factors

We evaluate whether the explanations reflect the attention of the CNN towards clinically relevant factors. This alignment with clinical factors is quantified as the IoU of the heatmaps with the segmentation masks of functionally different nuclei. The IoU is computed for 100 testing images of the PanNuke dataset containing at least one neoplastic nucleus (indicative of the presence of tumor). The heatmaps are thresholded, as in (Zhou et al. 2018), so that they activate on average for 60% of the pixels of the positive class images. We obtain one IoU score per image and per annotation type. Because some nuclei types are not present on some subsets of images, the IoU for a given annotation type is computed only on the subset of images that contains at least one instance of this type. Results are presented in Fig. 3. The IoU of the heatmaps generated for a CNN with fully randomized weights is added as a baseline for comparison[4].

Consistency, Repeatability and Randomization Test

We evaluate the consistency of the visualizations by assessing three points: (i) their dependency on the XAI hyper-parameters, (ii) their repeatability, and (iii) their response to the randomization test. Since activation maps do not require the tuning of hyper-parameters, the consistency (i) and repeatability analysis (ii) are only reported for LIME.

To assess (i), we monitor the changes in the SSIM over small shifts in two hyper-parameters, namely the number of samples (that corresponds to the neighborhood size), and the number of features, i.e. the number of superpixels retained for the analysis. In (i), differently from the analysis in (ii), we do not reset the initialization seed used by the LIME surrogate model across repetitions. This gives more stability to the comparison, reducing the level of stochasticity within the generation of LIME explanations. The plot in Fig. 4a shows the SSIM for changing values of the neighborhood size, with the number of superpixels fixed at 100. The shift in the hyper-parameters was performed within the range of zero to 3000 with a step of 50. For a given value N on the x-axis, the plot represents the SSIM between the heatmap obtained with N samples (image perturbations) and the one obtained with $N - 50$ samples. Similarly, Fig. 4b reports the SSIMs between heatmaps obtained by fixing the number of samples to 1000 and shifting the number of superpixels retained from zero to 3000 by intervals of 50.

[4]This is not the same as the cascaded randomization test proposed in (Adebayo et al. 2018), that is reported in Fig. 7

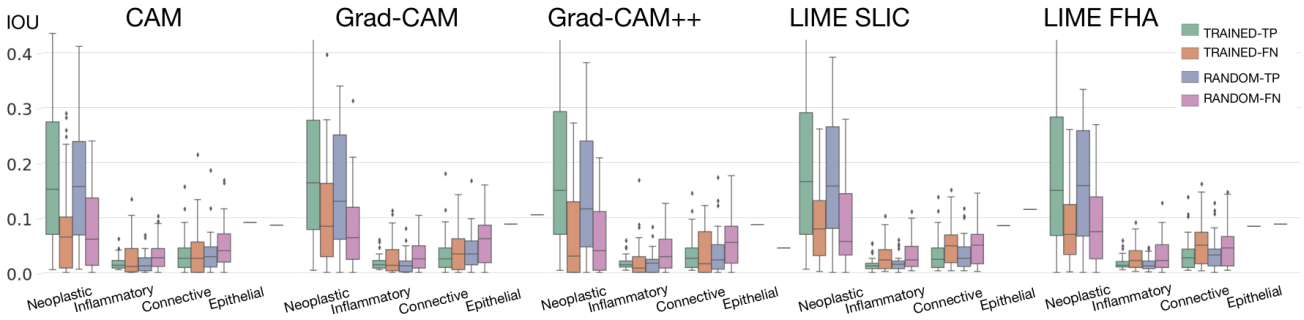
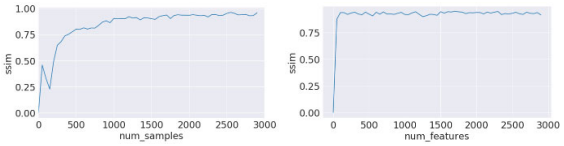
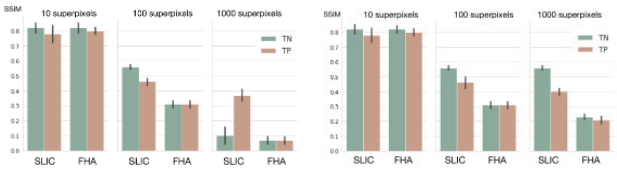


Figure 3: Quantification of CNN attention on the nuclei types in PanNuke expressed as the IoU in the testing set. The IoU of a network with randomly initialized weights (RANDOM-TP and RANDOM-FN) is added as a baseline for comparison. Best seen on screen.



(a) Neighborhood size (b) Number of superpixels

Figure 4: SSIM between heatmaps obtained from LIME when a parameter differs by a shift of 50. The studied parameter is the number of samples in (a) and the number of superpixels in (b). E.g. the SSIM for the x-axis point 1000 is the SSIM between the LIME method with 1000 samples and the LIME method with 950 samples. The number of superpixels is set to 100 in (a) and the neighborhood size to 1000 in (b).



(a) Neighborhood size of 100 samples. (b) Neighborhood size of 1000 samples.

Figure 5: SSIM evaluating LIME repeatability over 25 repetitions for True Positive (TP) and False Negative (FN) inputs for both SLIC and FHA superpixels. The random seed initialization for LIME is reset at every repetition. Heatmaps obtained with 10, 100 and 1000 superpixels are compared. Error bars report the standard deviation.

We assess (ii), namely the repeatability of LIME visualizations, by evaluating the SSIM of the heatmaps obtained with 25 different initialization seeds. As the hyper-parameter values for the number of superpixels and the neighborhood size may also affect LIME heatmaps (see Fig. 4), we compare in Fig. 5 the repeatability of the visualizations for 10, 100 and 1000 superpixels with neighborhoods of 100 and 1000 samples. High repeatability (SSIM around 0.8) is obtained only with 10 superpixels and a larger neighborhood Fig. 5(b) results in slightly more stable explanations.

The randomization test (iii) is finally performed for the XAI visualization methods considered in this analysis to assess the dependency of the explanations on the CNN weights. The weights are randomized in a cascading fashion (from the top to the bottom layers) up to the full randomization of all the CNN. In Fig. 6, we show visual examples of these cascaded randomizations. The SSIM between the original heatmap (from the trained CNN) and the one after the randomization at each layer is shown in Fig. 7.

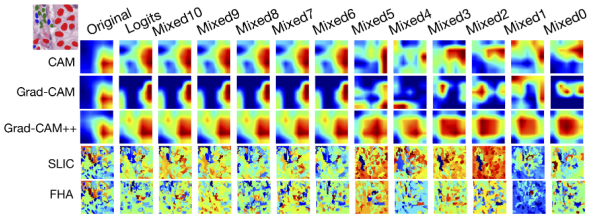


Figure 6: Visual assessment of the cascading randomization (Adebayo et al. 2018) test for the positive input shown in the top left corner. The progression from left to right shows the XAI heatmaps for a randomization of the network weights up to the layer represented by each column.

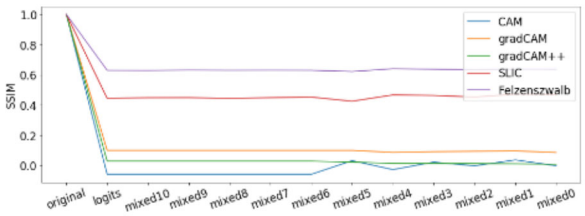


Figure 7: SSIM for cascading randomization (Adebayo et al. 2018). We show the SSIM between the original heatmaps (from the trained network) and the heatmaps generated as the CNN weights are randomized in the cascading way.

Discussion

The experiments propose an in-depth evaluation of popular XAI methods. XAI explanations were generated and visually compared for a total of 200 testing inputs, despite only a subset is presented in Fig. 1. The full set of results can be inspected in the GitHub repository. The CAM-based visualizations in Fig. 1 seem similar to each other. They all activate more frequently and with larger absolute values for positive predictions, with stronger intensity on the neoplastic nuclei areas, as also seen in (Graziani, Andrearczyk, and Müller 2019). The results obtained with LIME, on the other hand, are difficult to interpret in the reported examples, being very dependent on the generation of superpixels (Palatnik de Sousa, Maria Bernardes Rebuzzi Vellasco, and Costa da Silva 2019). Apart from these considerations, however, the simple qualitative analysis is not sufficient to evaluate the appropriateness and reliability of these XAI methods for histopathology images.

The quantitative metrics in this paper evaluate XAI methods from a global perspective. The correlation analysis in Fig. 2a shows that the explanations mostly agree when the prediction is low (i.e. low probability of tumor). The agreement is given by the very low values everywhere on the heatmap for these inputs (as also seen in the qualitative assessment in Fig. 1). The SSIM values in Fig. 2 further show that XAI visualizations are dissimilar for positive predictions (TPs and FPs). The similarity between CAM and Grad-CAM, shown by the high SSIM values for all prediction outcomes (0.8 on average), confirms the equivalence between the two methods explained in (Selvaraju et al. 2017).

The IoU analysis in Fig. 3 suggests that all XAI methods attribute the largest attention to the neoplastic nuclei. At first sight, this may indicate that neoplastic nuclei are responsible for the identification of tumorous patches. There is, however, an important bias in the distribution of neoplastic nuclei, that are in larger numbers than all other types and are present only in positive images by construction (the ground truth of all patches is computed by looking at the neoplastic nuclei). We cannot thus clearly establish the correlation between the attention attributed to neoplastic nuclei and the network outcome. Because of this, we compare the IoUs of a trained CNN to that of a randomly initialized CNN, also reported in Fig. 3. Only a slight (non-statistically significative) reduction in the IoU with the neoplastic nuclei is noticed. This suggests that the likelihood of high IoU with neoplastic nuclei is already high for a random heatmap and that the alignment of the XAI visualization with clinically relevant features is only apparent.

The consistency of the explanations upon small shifts in the hyper-parameter setting is another point of our analysis. Fig. 4 shows that LIME results depend strongly on the hyper-parameter set-up. The SSIM evaluating shifts in the parameters only plateaus after passing high values for both the neighborhood size (above 1000 samples) and the number of features (i.e. number of superpixels, above than 100). The latter should be much larger than general recommendations for LIME on non-visual inputs (e.g. around 10 features for text applications). Besides, Fig. 5 shows that LIME explanations are very unstable and not repeatable unless the number of features used for the analysis is small (around 10, from Fig. 5). The two analyses in Fig. 4b and 5 show that it is not possible to obtain both repeatable and consistent LIME visualizations. Heatmaps with only 10 superpixels, moreover, are hard to visually interpret (not illustrated in this paper, but available on the GitHub for comparison).

We observe in Fig. 6 that XAI visualizations change when the network weights are randomized at cascading depths. Especially the results for CAM-based methods align with those in (Adebayo et al. 2018), showing diffused heatmaps around the image center. The cascading randomization test is passed by CAM-based methods in Fig. 7.This does not mean, however, that the CAM-based visualizations of the trained CNN point more towards clinically relevant regions than an untrained CNN. As previously mentioned, the comparison of the IoUs of trained and untrained CNNs in Fig. 3 shows that XAI visualizations for the trained CNN do not overlap more with neoplastic nuclei than those for an untrained CNN. We conclude that the XAI visualizations analyzed in this paper reflect only apparently the clinical variability in the images.

Conclusions

This study proposes qualitative and quantitative analyses of XAI visualization methods in the context of CNNs for digital pathology. The qualitative inspection may lead to misleading conclusions, e.g. that the network's attention points towards neoplastic nuclei. As neoplasticity is a main indicator of tumor, these explanations increase the confirmation bias and the tendency of accepting the CNN decisions as true. Our evaluation however, shows no significant difference in the attention paid to neoplastic nuclei by a trained CNN than that of a randomly initialized one. The explanations, therefore, do not seem to increase their alignment with clinical evidence after network training. In addition, LIME visualizations are neither consistent nor repeatable, generating different explanations for different settings of the hyper-parameters or initialization seeds. Driven by these results, we feel that XAI visualizations should not be used as a way of proving the correctness of the CNN decision-making on this task.

We remark a limitation of this study in the annotations of nuclei types, that are not exhaustive of the clinical factors that could be involved in the decisions. Mitoses, for example, are not considered by our analyses.We focus on the most popular XAI methods, not proposing an exhaustive coverage. Our framework for the evaluation, however, can be easily extended to new annotations, architectures, datasets, and XAI methods. This paper shows that more applied research is indeed crucial to test and confirm the utility of XAI methods in the medical imaging field.

Acknowledgments

This work is supported by the European Union's projects PROCESS (agreement n. 777533), ExaMode (n. 825292) and AI4Media (n. 951911). NVIDIA Corporation supported this work with the donation of the Titan X GPU.

References

Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; and Süsstrunk, S. 2012. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence* 34(11):2274–2282.

Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; and Kim, B. 2018. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, 9505–9515.

Arun, N.; Gaw, N.; Singh, P.; Chang, K.; Aggarwal, M.; Chen, B.; Hoebel, K.; Gupta, S.; Patel, J.; Gidwani, M.; et al. 2020. Assessing the (un) trustworthiness of saliency maps for localizing abnormalities in medical imaging. *arXiv preprint arXiv:2008.02766*.

Chattopadhay, A.; Sarkar, A.; Howlader, P.; and Balasubramanian, V. N. 2018. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 839–847. IEEE.

Felzenszwalb, P. F., and Huttenlocher, D. P. 2004. Efficient graph-based image segmentation. *International journal of computer vision* 59(2):167–181.

Gamper, J.; Koohbanani, N. A.; Benet, K.; Khuram, A.; and Rajpoot, N. 2019. PanNuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In *European Congress on Digital Pathology*, 11–19. Springer.

Gamper, J.; Koohbanani, N. A.; Graham, S.; Jahanifar, M.; Khurram, S. A.; Azam, A.; Hewitt, K.; and Rajpoot, N. 2020. Pannuke dataset extension, insights and baselines. *arXiv preprint arXiv:2003.10778*.

Graziani, M.; Andrearczyk, V.; and Müller, H. 2018. Regression concept vectors for bidirectional explanations in histopathology. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*. Springer. 124–132.

Graziani, M.; Andrearczyk, V.; and Müller, H. 2019. Visualizing and interpreting feature reuse of pretrained cnns for histopathology. In *IMVIP 2019: Irish Machine Vision and Image Processing Conference Proceedings*. Irish Pattern Recognition and Classification Society.

Graziani, M.; Andrearczyk, V.; Marchand-Maillet, S.; and Müller, H. 2020. Concept attribution: Explaining CNN decisions to physicians. *Computers in Biology and Medicine* 123:103865.

Khan, A.; Atzori, M.; Otálora, S.; Andrearczyk, V.; and Müller, H. 2020. Generalizing convolution neural networks on stain color heterogeneous data for computational pathology. In *Medical Imaging 2020: Digital Pathology*, volume 11320, 113200R. International Society for Optics and Photonics.

Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C. J.; Wexler, J.; Viégas, F.; and Sayres, R. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *ICML*.

Kindermans, P.-J.; Hooker, S.; Adebayo, J.; Alber, M.; Schütt, K. T.; Dähne, S.; Erhan, D.; and Kim, B. 2019. The (un) reliability of saliency methods. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer International Publishing*.

Litjens, G.; Bandi, P.; Ehteshami Bejnordi, B.; Geessink, O.; Balkenhol, M.; Bult, P.; Halilovic, A.; Hermsen, M.; van de Loo, R.; Vogels, R.; et al. 2018. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *GigaScience* 7(6):giy065.

Lombrozo, T. 2006. The structure and function of explanations. *Trends in cognitive sciences* 10(10):464–470.

Madhyastha, P., and Jain, R. 2019. On model stability as a function of random seed. *arXiv preprint arXiv:1909.10447*.

Palatnik de Sousa, I.; Maria Bernardes Rebuzzi Vellasco, M.; and Costa da Silva, E. 2019. Local interpretable model-agnostic explanations for classification of lymph node metastases. *Sensors* 19(13):2969.

Rakha, E. A.; El-Sayed, M. E.; Lee, A. H.; Elston, C. W.; Grainge, M. J.; Hodi, Z.; Blamey, R. W.; and Ellis, I. O. 2008. Prognostic significance of nottingham histologic grade in invasive breast carcinoma. *Journal of clinical oncology* 26(19):3153–3158.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Tonekaboni, S.; Joshi, S.; McCradden, M. D.; and Goldenberg, A. 2019. What clinicians want: contextualizing explainable machine learning for clinical end use. *arXiv preprint arXiv:1905.05134*.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929.

Zhou, B.; Bau, D.; Oliva, A.; and Torralba, A. 2018. Interpreting deep visual representations via network dissection. *IEEE transactions on pattern analysis and machine intelligence* 41(9):2131–2145.