

Decision validity of heatmap-based explainable neural networks – an explanation first approach

G. Galiger

Abstract—The applicability of image processing models in digital histopathology highly depends on the ability of explaining model decisions. In this paper, we present a quantitative approach for measuring the validity of heatmap-based neural network decisions by defining two decision validity scores. The proposed validation method is benchmarked on Bag-of-local-Features models trained for detecting different tissue types in whole-slide images. Using the introduced validity scores, we are able to quantify the effects of different training factors on model decision-making, such as the size of the training dataset, or the usage of data augmentation. Moreover, we present an explanation first approach by leveraging slight performance trade-offs for considerably improving the understandability of explanations.

Index Terms—decision validity, explainable neural network, post-hoc explanation, whole-slide image, tissue type detection

I. INTRODUCTION

The appearance of high-resolution whole-slide images (WSI) in the field of histopathology enabled the usage of image processing techniques for the digital analysis of biopsies on a cellular level. Consequently, various computer-aided diagnosis (CAD) systems were developed for the analysis of WSIs to provide assistance for pathology experts. Despite black-box neural networks showing high image processing performance, the inability to explain model decisions greatly hinders the applicability of such approaches in all medical fields [6]. Therefore, explainability is an essential requirement of CAD systems intended for practical applications.

Due to task-related complexities, some CAD approaches might leverage a-priori information provided by human experts. As a related example, the tissue segmentation method proposed in [4] fully relies on radiologists initially identifying tissue types in input images. In [3] an explainable approach was introduced to automate this process by solving the initial task of tissue type detection using explainable CNNs based on the BagNet architecture [2]. While the detection performance of these models was thoroughly analyzed here, the analysis of model decisions was limited to a qualitative analysis. The main aim of this paper is to provide a formal method of quantifying the decision validity of these models, while also discussing detection performance aspects of the presented explanation first approach. Throughout the paper, we refer to the validity of model decisions as the correspondence of model activations with the expert-annotated ground truth.

II. METHODS

Fully automating the segmentation method in [4] requires the simultaneous detection of four tissue types. We split this

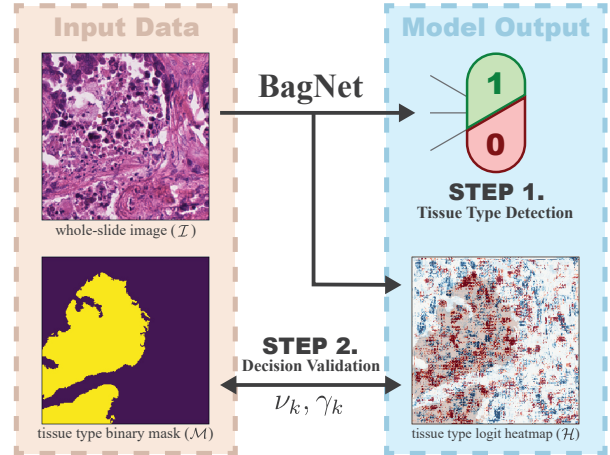


Fig. 1: Proposed tissue type detection process diagram.

task into single binary classification problems for each tissue type with the following classes: 1 – if the tissue type is detected in the input WSI, 0 – otherwise. This separation of classes contributes to making the analysis of decisions more easily understandable, as it eliminates a major complexity of simultaneous detection, namely having to explain relations between different tissue types. As later shown in Section III-A, this explanation first reformulation does not cause a significant loss in detection performance. As illustrated in Figure 1, the process of detecting one specific tissue type in a WSI patch is the first step of the proposed method.

A. Dataset adaptation

The data used for training and validating the BagNet models originates from [4], where two weakly-annotated patch-based WSI datasets were published, with class-level labels of four tissue types for all images, and expert-annotated pixel-level segmentation masks for testing and validation subsets. These datasets were transformed in eight (2 orig. datasets $\times$ 4 tissue types) binary classification datasets, adapting both the labels and masks to the binary problem. The class-level labels were utilized for training and testing our detection models, while pixel-level masks were used for validating model decision-making. In addition, data augmentation techniques were also used on all datasets to facilitate model generalizability.

B. Tissue type detection with BagNet models

Explainability characteristics are key factors when choosing models for real-world medical applications. Based on our

explanation first approach, the ease of understanding decisions was the main criterion for model choice. Hence, we adapted the BagNet architecture [2] for the newly defined binary classification problems. As opposed to similar heatmap-based networks (e.g. Grad-CAM), BagNet decisions are based on image patches that are independent of each other, thus making explanations more intuitive to define and understand.

The adaptation of the architecture consisted in changing the output size of the linear classifier and replacing the softmax operator with sigmoid, as the binary classification problems assume class independence by definition, i.e. a specific tissue type can either be found in the input image, or not. The ResNet-50 backbone of the original BagNet architecture was kept, while transfer learning was also leveraged in all further experiments using ImageNet pre-trained weights. Further implementation details can be found in [3].

C. Quantitative measurement of decision validity

To further increase trust towards such approaches, the validity of model decision-making should also be analyzed as a second step after the actual tissue type detection, as illustrated in Figure 1. This can be done by comparing the available pixel-level binary masks annotated by pathologists with the BagNet logit heatmaps associated with the respective input images.

To introduce a quantitative metric for this validation, we first formalize an input image as

$$\mathcal{I} : \mathcal{U} \longrightarrow [0, 1]^c,$$

where $\mathcal{U} = [[0; p - 1] \times [0; q - 1]]$ is the set of all pixel index pairs in the image, p and q are row and column dimensions, and c is the number of color channels. Consequently, $\mathcal{I}(i)$ represents the value of the i -th pixel of our image, e.g. for $i = (0, 0)$, $\mathcal{I}(i)$ returns the color channel values of the top-left corner pixel. For the present application of tissue type detection RGB images of 224×224 pixels are used as input. Hence, the parameters of the previous definition are $c = 3$ and $p = q = 224$. Note that in most software implementations color channel intensity is represented with a discrete number, usually using the set $\{0, 1, \dots, 255\}$ instead of $[0, 1]$.

From this formulation, the binary mask and the BagNet heatmap for a tissue type in an image can be written as

$$\mathcal{M} : \mathcal{U} \longrightarrow \{0, 1\}, \text{ and } \mathcal{H} : \mathcal{U} \longrightarrow \mathbb{R}$$

respectively. Note that this notation of $\mathcal{H}$ is not limited to BagNet, as it can represent any logit heatmap obtained using similar approaches, e.g. layer-wise relevance propagation [1].

Finally, we note $\mathcal{U}_k \subset \mathcal{U}$, $k \in (0, 100]$ as the set of top $k\%$ pixel indexes from $\mathcal{U}$ with the largest $\mathcal{H}(i)$, $i \in \mathcal{U}$ heatmap values, and $\mathcal{U}_{\mathcal{M}} \subset \mathcal{U}$ the relevant pixel indexes based on the binary mask, i.e. where $\mathcal{M}(i) = 1$. Using these notations, we can define the decision validity score ν_k as

$$\nu_k = \frac{|\mathcal{U}_k \cap \mathcal{U}_{\mathcal{M}}|}{|\mathcal{U}|} = \frac{|\mathcal{U}_k \cap \mathcal{U}_{\mathcal{M}}|}{p \cdot q} \in [0, 1].$$

Note that using this notation $\mathcal{U}_{100} = \mathcal{U}$, so $\nu_{100} = \frac{|\mathcal{U}_{\mathcal{M}}|}{p \cdot q}$ returns the mask – input image size ratio in pixels, which can be used to define a derived validity score

$$\gamma_k = \frac{\nu_k}{\nu_{100}} \in [0, \infty),$$

TABLE I: Comparing BagNets and non-explainable models.

Model	Acc.	Prec.	Recall	Spec.
LUAD – LYM				
BagNet-9	0.908	0.9225	0.8983	0.8581
BagNet-17	0.954	0.9175	0.9783	0.966
BagNet-33	0.873	0.7675	0.9433	0.9003
LUAD – NEC				
VGG19	0.925	0.7692	1.0	0.9
ResNet50	0.95	0.8333	1.0	0.9333
BagNet-17	0.975	0.9091	1.0	0.9667

that also accounts for the mask size relative to the size of the input image. The γ_k score can be seen as a generalization of the feature portion (FP) metric used for similar purposes of plausibility analysis in medical applications [5].

In practice, a ν_k score close to 1 suggests a valid decision, as a $\nu_k = 1$ score indicates a complete correspondence to the ground truth. In contrast, quantifying the randomness of a decision with this score is dependent on the mask size, i.e. $\nu_k \leq \nu_{100}$, the value of ν_{100} indicating the difficulty of obtaining a valid decision. Hence, a non-valid decision is better illustrated with the derived γ_k score, which indicates a random decision when close to 1.

III. RESULTS

A. BagNet architecture for tissue type detection

The performance of BagNet models in certain applications is dependent on the receptive field size of the chosen architecture. To choose the right model for the present application, all three originally proposed BagNet architectures were evaluated on multiple binary datasets. Table I shows a representative example of this experiment's results on the LUAD – LYM dataset, which concluded choosing the BagNet-17 model for further experiments.

As using explainable models generally entails performance sacrifices [2], we explored how the adapted BagNet compares to other non-explainable models. We performed a comparison between BagNet-17 and both ResNet-50 and VGG-19 architectures, the results of which are shown in Table I on the LUAD – NEC dataset [3]. This experiment showed that the performance of these three models are dataset-dependent and relatively similar. Moreover, the explainable model even outperforms the others in some cases, as shown in this example.

B. Explainability analysis and validity of decisions

Heatmaps offer qualitative visual representation of the decision-making process, however, quantitative measures like ν_k and γ_k are also needed for a comprehensive analysis of model decisions. Both scores measure the amount of important pixels that coincide with the expert-annotated binary masks, which can be viewed as the ground truth for decision-making of the present application. As a result, higher validity scores indicate better adherence to the expected decision-making process, i.e. that of expert pathologists. For conciseness, we perform a systematic analysis of BagNet decisions in this

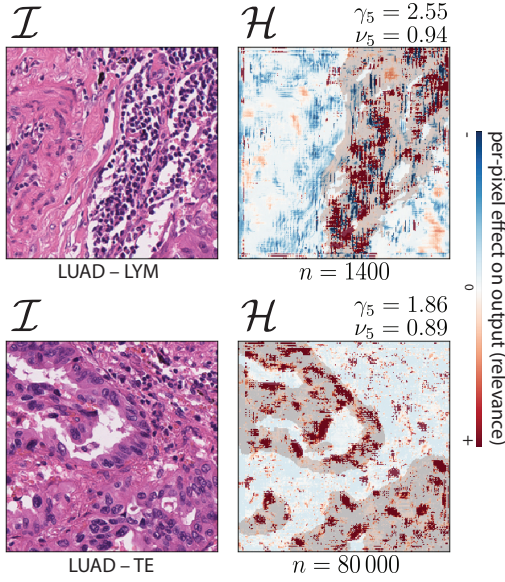


Fig. 2: Effects of training data size (n) on model decisions. Shown on heatmap ($\mathcal{H}$) activations of two WSIs ($\mathcal{I}$) from datasets with different sizes, and respective ν_5 , γ_5 scores.

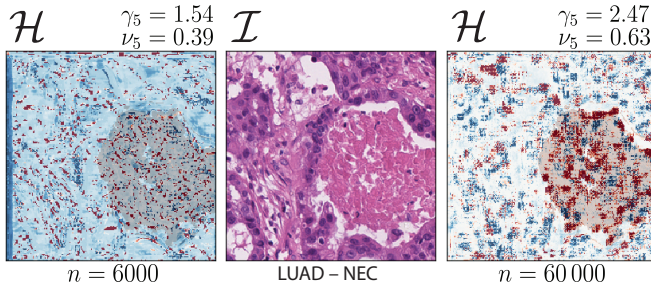


Fig. 3: Effects of data augmentation on decision-making.

section based on both scores with $k = 5$, a representative value determined empirically.

To optimize the training process of the adapted BagNet model for the detection of tissue types, we first evaluate the effects of training data size (n) on model decisions. Two representative examples for this experiment are shown in Figure 2, which demonstrate the low impact of n on the BagNet decision-making process, as ν_5 and γ_5 scores indicate high decision validity for both samples. This observation corresponds to previous experiments, which concluded tissue-related textural differences having high influence on model performance, rather than the size of training datasets [3].

Another important aspect of optimizing the BagNet training process that is quantifiable by the ν_5 and γ_5 scores is the effect of data augmentation on model decisions. This is illustrated in Figure 3 with the heatmaps of the same input image, where the network was trained with (right) and without data augmentation (left). As shown here, the model is not able to learn the detection task without data augmentation in case of some tissue types. This is visible on the low ν_5 and close-to-1 γ_5 scores of the left-side example, where the heatmap also indicates an almost random decision.

In addition, this figure also demonstrates the necessity for

using both validation scores for the quantitative analysis of decision-making. Although both ν_5 and γ_5 values are low for the left-side sample, using one score to quantify the decision validity for the right side is not trivial. In spite of the significant improvement on the augmented dataset, ν_5 is not particularly high because of the difficulty of obtaining a valid decision for this example, i.e. the small relative mask size ν_{100} . However, this is accounted for with γ_5 , which indicates similar decision validity to the examples from Figure 2.

IV. DISCUSSION

The results presented in this paper demonstrate the applicability of explanation first CAD approaches in the field of digital pathology. The heatmap-based BagNet model shows better performance than similar non-explainable neural networks in some applications, while also being able to provide a human-readable explanation of decisions. We argue that ensuring the applicability of such methods in medical fields also requires performing comprehensive studies of decision validity. Therefore, we defined two quantitative metrics, which we used to analyze and compare the effects of different training strategies on model decisions.

To simplify the analysis process, we avoided differentiating single heatmap pixels from each other in both decision validity scores. However, individual weights could also be assigned to pixels based on their heatmap values to better quantify decision certainty. Moreover, analyzing pixels with the highest negative influences on model decision (i.e. lowest negative $\mathcal{H}(i)$, $i \in \mathcal{U}$ values) could also yield valuable quantitative information about model decision-making. In addition, we would also like to extend the validity analysis on entire datasets, and analyze further aspects of the training process using the proposed scores, e.g. changes in network architecture, training parameters, or using data from different sources.

ACKNOWLEDGMENT

Supported by the ÚNKP-23-3 New National Excellence Program of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund. Part of the research was supported by Babeş-Bolyai University (No. 36586/25.11.2022), the Márton Áron College of ELTE, and the Collegium Talentum Programme of Hungary.

REFERENCES

- [1] S. Bach, A. Binder, G. Montavon, F. Klauschen, K. R. Müller, and W. Samek, *On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation*. PloS one, 10(7), p.e0130140, 2015.
- [2] W. Brendel and M. Bethge, *Approximating cnns with bag-of-local-features models works surprisingly well on imagenet*. arXiv preprint arXiv:1904.00760, 2019.
- [3] G. Galiger and Z. Bodó, *Explainable patch-level histopathology tissue type detection with bag-of-local-features models and data augmentation*. Acta Universitatis Sapientiae, Informatica, 15(1), pp.60-80, 2023.
- [4] C. Han, J. Lin, J. Mai, Y. Wang, Q. Zhang, B. Zhao, X. Chen, X. Pan, Z. Shi, Z. Xu and S. Yao, *Multi-layer pseudo-supervision for histopathology tissue semantic segmentation using patch-level classification labels*. Medical Image Analysis, 80, p.102487, 2022.
- [5] W. Jin, X. Li, M. Fatehi and G. Hamarneh, *Guidelines and evaluation of clinical explainable AI in medical image analysis*. Medical Image Analysis, 84, p.102684, 2023.
- [6] B. H. Van der Velden, H. J. Kuijff, K. G. Gilhuijs and M. A. Viergever, *Explainable artificial intelligence (XAI) in deep learning-based medical image analysis*. Medical Image Analysis, 79, p.102470, 2022.