Acta Univ. Sapientiae Informatica **15,** 1 (2023) 60–80

DOI: 10.2478/ausi-2023-0006

Explainable patch-level histopathology tissue type detection with bag-of-local-features models and data augmentation

Gergő GALIGER
Babeş–Bolyai University
Cluj-Napoca, Romania
email: galigergergo@yahoo.com

Zalán BODÓ
Babeş–Bolyai University
Cluj-Napoca, Romania
email: zalan.bodo@ubbcluj.ro

Abstract. Automatic detection of tissue types on whole-slide images (WSI) is an important task in computational histopathology that can be solved with convolutional neural networks (CNN) with high accuracy. However, the black-box nature of CNNs rightfully raises concerns about using them for this task. In this paper, we reformulate the task of tissue type detection to multiple binary classification problems to simplify the justification of model decisions. We propose an adapted Bag-of-local-Features interpretable CNN for solving this problem, which we train on eight newly introduced binary tissue classification datasets. The performance of the model is evaluated simultaneously with its decision-making process using logit heatmaps. Our model achieves better performance than its non-interpretable counterparts, while also being able to provide human-readable justification for decisions. Furthermore, the problem of data scarcity in computational histopathology is accounted for by using data augmentation techniques to improve both the performance and even the validity of model decisions. The source code and binary datasets can be accessed at: https://github.com/galigergergo/BolFTissueDetect.

Key words and phrases: computational pathology, tissue type detection, explainable artificial intelligence, convolutional neural networks, data augmentation

1 Introduction

The accurate interpretation and analysis of medical images can aid in the diagnosis of various diseases and conditions [9], provide information about the progression of diseases [24], and assist in treatment planning [25]. In the field of histopathological image processing, digital images are obtained as a result of the invasive technique of removing and scanning specific tissue samples from a patient's body through some form of biopsy [8]. The scanned whole-slide images (WSI) represent high-resolution images of gigapixel order, which results in a very high computational cost of analyzing them with different algorithms [16]. To solve the problems caused by WSI sizes, single whole-slide images can be broken down into several smaller patches, and then processed in adequately sized batches depending on available resources.

The introduction of WSIs made it possible to develop different computer-aided diagnosis (CAD) and prognosis systems for automating various tasks in the field of medical image processing [15]. The task in the field of WSI processing examined in this paper is cancer detection, more specifically the detection of different tissue types in WSIs of tumor microenvironments (TME). TME has a significant impact on tumor initiation and progression [12] and also affects the prognosis and response to therapy of cancer patients [1].

Automating different tasks of trained human pathologists using CAD systems poses a variety of different challenges by itself. For instance, machine learning-based methods may be vulnerable to only changing a single pixel of an image [19]. In case of WSIs different artifacts (air bubbles, dirt) may contaminate tissue samples, and different forms of deterioration (tears, cracks, color variations) of the samples are also common [15], which could all bias machine-learning models. Furthermore, creating a large, high-quality annotated medical dataset, which is essential for all tasks involving supervised learning, is also considered a difficult task. Correct annotation of WSIs is a time-consuming, costly process that requires the involvement of a large number of trained professionals [2]. Moreover, privacy concerns may arise when working with specific medical conditions [20]. Another major challenge of using CAD systems in the medical sector arises from the black-box nature of the current state-of-the-art approaches [22]. Although deep learning methods have shown promising performances in medical tasks, the inability to explain decisions has raised concern among medical experts [21].

Due to task-related complexities, histopathological segmentation methods might rely on additional, a-priori information besides the input images. As a related example, in [11], the proposed state-of-the-art TME tissue segmenta-

tion method integrates the usage of classification labels of the input images to achieve good performance. However, no method of identifying these classification labels was introduced in this paper, the authors implying the task to be carried out by human radiologists. Thus, the proposed method represents a semi-automated approach, which heavily relies on human pathologists identifying different tissue types in WSIs.

The main focus of this paper is to fully automate this process by introducing a way of identifying the TME tissue types used as inputs for the segmentation method proposed in [11]. As this approach intends to replace human experts, the reliability of its decision-making process is of utmost importance. Therefore, we obtain highly interpretable problems by reformulating the task of TME tissue type detection as multiple binary classification problems. We refer to the interpretability of model decisions as the correspondence of the model activations and the ground truth. For solving the task of binary tissue classification, we propose an adapted version of the explainable BagNet architecture [5], which demonstrates matching performance compared to human experts. Moreover, the usage of the BagNet architecture also provides a solution for the concerns regarding the black-box nature of models in the medical field by being able to justify decisions taken. The proposed models show a decision-making process that aligns with the expected histopathological reasoning of human radiologists. In addition, we solve the problem of medical data scarcity with data augmentation techniques, the effects of which on both the performance and validity of model decisions are further analyzed in detail.

Overall, we present four main contributions in this paper: (i) The task of TME tissue type detection is reformulated to multiple binary classification tasks for the sake of clear, human-readable justification of model decision-making. (ii) The BagNet Bag-of-local-Features explainable CNN architecture is adapted for this task and the validity of its decisions is evaluated using logit heatmaps. (iii) Eight binary datasets are created for this task from two publicly available ones and the effects of expanding them using data augmentation techniques are also reviewed. (iv) A detailed quantitative and qualitative evaluation of the adapted method is performed on the newly created datasets.

The rest of this paper is structured as follows: Section 2 describes the specific TME binary classification task formulation and necessary adaptations in both model architecture and dataset structure. In Section 3, we present the experiments used to benchmark the adapted BagNet model on the binary classification task, along with an in-depth analysis of obtained results. We conclude our findings in Section 4, where we also provide further research directions regarding this topic.

2 Methodology

2.1 Problem formulation

For the task of identifying patch-level labels, such as in the case of the training samples from the two presented datasets, individual patches have to be assigned to four different classes, thus resulting in a multi-class multi-label classification task. Successfully solving this problem, while also providing reliable explanations for decisions is highly difficult in the case of such a complex classification task. Therefore, the task should be reformulated to solve one binary classification task for each of the four classes, i.e. detecting whether or not one type of tissue (class) is present in a certain patch. This reformulation yields multiple significantly simpler detection problems, which in turn simplify the interpretation process of explainable models used for solving these.

Our choice for solving these binary classification problems was the BagNet interpretable CNN architecture, which showed promising results in a variety of different tasks [5, 17, 18] and is inherently explainable with pixel-level activation heatmaps. As we are using weakly-annotated datasets with fully segmented validation and testing data, the exact location of tissue types in these cases is precisely known. This implies the natural application of heatmaps for decision interpretation since pixel activation validity can be verified using segmentation data.

BagNets were originally proposed for single-label multi-class classification tasks on the ImageNet dataset [7], which makes adapting it to binary classification problems a necessary procedure. This was done by changing the output size of the linear classifier following the average pooling operator to 1 and replacing the softmax operator with sigmoid to infer the image-level class evidence. By introducing the sigmoid operator, the independence of different classes is assumed, which is inherent in the case of the two classes of this binary classification problem: a certain tissue type is present in a patch versus it is not present.

2.2 Datasets

The scarcity of suitable quality datasets for medical image processing applications applies to WSIs of tumor microenvironments, which requires radiology expertise to be correctly annotated [2], and the publication of such datasets also raises concerns regarding patient privacy [20]. The datasets used in this paper originate from the previously mentioned paper [11], where two different

Binary dataset	Training	Testing	Validation
LUAD – LYM	1400	200	100
LUAD – NEC	6000	120	40
LUAD – TE	8000	80	60
LUAD – TAS	8000	140	40
BCSS – NEC	6000	800	800
BCSS – LYM	12 000	4000	2000
BCSS – STR	16 000	4000	1600
BCSS – TUM	18 000	4000	3000

Table 1: Size distribution of the 8 binary datasets between training, testing, and validation subsets.

weakly-annotated datasets of TMEs were created and published for further research:

LUAD-HistoSeg (LUAD): This dataset was specifically created for training the proposed segmentation model and included patches with four tissue categories: tumor epithelial (TE), tumor-associated stroma (TAS), necrosis (NEC) and lymphocyte (LYM).

BCSS-WSSS (BCSS): This dataset was adapted from a previously published fully-supervised dataset [2] and included four classes: tumor (TUM), stroma (STR), lymphocytic infiltrate (LYM), and necrosis (NEC).

In a similar manner to the used model architecture, the two original datasets would have been adequate for a different type of problem, a multi-class multi-label classification task. This resulted in both of the datasets having to be adapted to the newly formulated binary classification problems. The adaptation consisted in creating a binary classification dataset for each of the classes from the original datasets. This was done by first counting the positive and negative examples for a given class in a dataset, i.e. the number of samples that contained a given tissue type and the number of samples that did not. To obtain an optimal balance between the two classes, the binary datasets were constructed with the smaller one of the two binary classes (positive or negative) and a randomly sampled version of the larger binary class. Applying this method to all four classes of both datasets resulted in 8 (4×2) binary datasets of varying sizes. Table 1 shows the size distribution of the 8 binary datasets.

Transformation	Probability
Horizontal skew	1.0
Low-angle rotation	1.0
90-degree-multiple rotation	0.75
Horizontal flip	0.5
Vertical flip	0.5

Table 2: Probabilities of applying different data augmentation transformations when expanding binary datasets.

2.3 Data augmentation

The original LUAD and BCSS datasets are relatively small for training DNNs in a supervised manner, even for the task of binary classification. The problem of data scarcity is further accentuated by the fact that the actual datasets used for binary classification are even smaller than the original ones. This happens because these represent subsets of the original datasets obtained by their adaptation to the binary classification task, four binary datasets having been created from a single original dataset. To alleviate the issue of small datasets, we used data augmentation methods to expand all eight binary datasets to 10 times their original size.

To simulate common, practically occurring changes in WSIs such as different rotation and skew angles of slides during the scanning procedure, only related augmentation transformations were used. The augmentation techniques used for the expansion of the binary datasets were limited to morphological transformations, which included low-magnitude skew operations, random-angle rotations by a maximum of 3 degrees, rotations with 90-degree multiples, and flip operations. All of the listed transformations were applied with certain probabilities, as shown in Table 2. These values were chosen based on the following reasoning: skew and low-angle rotation are always applied to introduce some level of morphological diversity in every single new image, while the remaining three transformations further differentiate images, the probability for rotation being higher since flip operations are twice as many in number.

2.4 Implementation details

All the convolutional neural networks analyzed in our experiments were implemented in PyTorch and we used the Augmentor[1] Python package for aug-

[1]Augmentor Python Package (https://augmentor.readthedocs.io/en/stable)

menting the binary datasets. We trained the models using an NVIDIA Titan Xp GPU and utilized ResNet-50 as the classification backbone for the BagNet architecture. The resolution of the input patches was 224 × 224 and the batch size was set to 29, the maximum value we could obtain before reaching GPU memory limitations. The network weights were optimized for binary cross-entropy loss using stochastic gradient descent (SGD) with a weight decay of 0.0001 and a momentum of 0.9. The learning rate was set to 0.01 and divided by 10 every 30 training epochs. These hyperparameter settings were inspired by the ImageNet training example from the original PyTorch GitHub repository[2].

3 Experiments and results

3.1 Training process

3.1.1 Receptive field size

The main advantage of using the BagNet architecture lies in the simplicity of explaining its decision-making process, which is a consequence of its separation of receptive fields in inferring class evidence. From the originally proposed BagNet–q models, with $q \in \{9, 17, 33\}$ receptive field sizes, BagNet–17 showed the best classification performance on the ImageNet dataset [5]. However, the transferability of ImageNet results to WSI processing is questionable due to the inherent differences between natural images and tissue scans. As a result, we benchmarked all three BagNet models for binary tissue classification on the LUAD – LYM dataset to find the most suitable receptive field size for this application. The results of this experiment are shown in Table 3.

In contrast to the natural images from ImageNet, where the BagNet–17 model showed clearly superior classification performance, the results are less clear in this case. The smaller receptive field size of 9 × 9 pixels leads to the highest values in some of the calculated performance measures, although these are not significantly better than the ones obtained with 17 × 17 receptive fields, with below 2% differences in average validation precision, recall, and specificity. However, there are significant differences in performance in every other case where the values obtained by BagNet–17 are highest, mostly exceeding 5%, and even 10% in some cases. As a result of these findings, we use the BagNet–17 model in the following experiments, which we will also refer to as BagNet.

[2]PyTorch GitHub Repository (https://github.com/pytorch/examples/tree/main/imagenet)

Model	Acc.	Prec.	Recall	Spec.
Without data aug.				
BagNet–9	0.93	0.875	**0.9667**	**0.9459**
BagNet–17	**0.94**	**0.925**	0.95	0.925
BagNet–33	0.88	0.8	0.9333	0.8889
With data aug.				
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

Table 3: Results of training BagNet models with different receptive field sizes. The values shown in this table are the average validation accuracy, precision, recall, and specificity scores obtained after training the models on the LUAD – LYM dataset with and without using data augmentation.

3.1.2 Model backbone

Neural networks show a tendency to benefit from increasing depth in various image processing tasks [10]. In [5], the BagNet network was proposed with the ResNet-50 architecture as a backbone, which is considered a highly layered architecture by current standards [13]. In order to test the necessity of such a deep backbone architecture for the binary classification task of TME tissue types, we trained the BagNet model on the LUAD – LYM dataset using two different backbones: the previously mentioned ResNet-50, and a significantly smaller CNN, which consisted of two convolutional layers of 32 and 64 neurons respectively, followed by a ReLU activation layer.

To present the findings of this experiment, the average validation accuracy values of the two models are illustrated on the left side of Figure 1. Using the small CNN as a backbone, the accuracy values converge in around 30 epochs, not improving considerably on initial values. This could be caused by the small network's inability to sufficiently generalize the complex image processing problem. In contrast, the significantly deeper ResNet-50 backbone leads to later conversion but shows substantially improved classification performance. Therefore, we use the ResNet-50 as the backbone for our BagNet models in the following experiments.

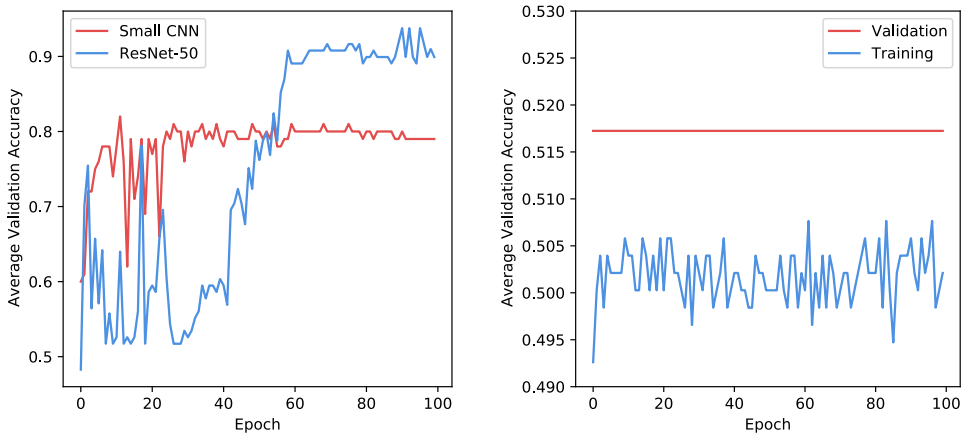

Figure 1: Results of benchmarking small CNN and ResNet-50 backbones for BagNet for binary classification trained on the LUAD – LYM dataset.

3.1.3 Transfer learning

Likewise to many other architectures, leveraging transfer learning, i.e. initializing BagNet weights with weights originally trained on the ImageNet dataset leads to significant performance increases in many different tasks [4, 17]. To verify the validity of the premise for the task of TME tissue binary classification, we evaluated three different ways of weight initialization for the adapted BagNet model:

Default initialization: No pre-trained weights were loaded in this case, initialization was done by the default PyTorch weight initialization process.

Loading ImageNet pre-trained weights: In this case, weights were loaded from the original pre-trained BagNet model published with the original paper [5]. As with the network architecture, adaptations had to be carried out in the case of the pre-trained weights as well, since the last linear transformation layer of the architectures did not match. The adaptation consisted in averaging the weights corresponding to the 2048 ImageNet classes in order to obtain a single weight tensor.

Loading LUAD – NEC pre-trained weights: We managed to obtain promising results by training the BagNet model with default initialization on the LUAD – NEC binary dataset. The weights of this pre-trained model were also used in this benchmark for initialization for training on binary datasets different from LUAD – NEC.

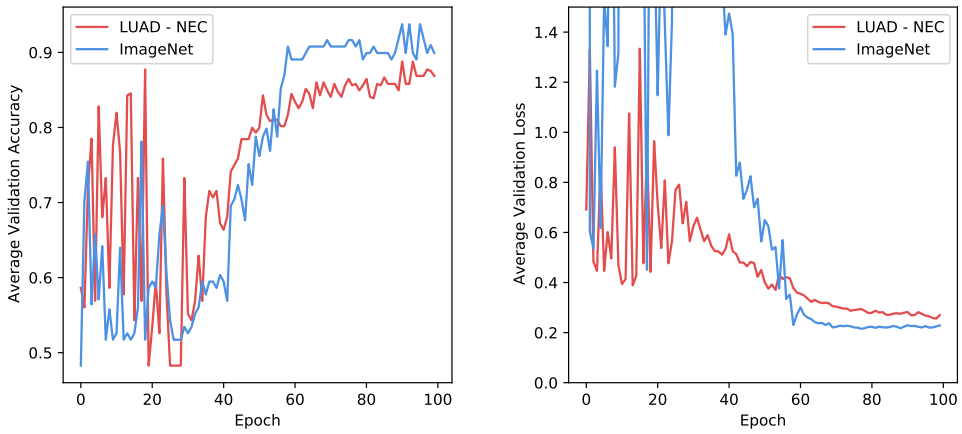

Figure 2: Results of training BagNet with ResNet-50 backbone on the LUAD – LYM dataset with weight initialization using pre-trained models on ImageNet and LUAD – NEC datasets.

These weight initialization approaches were benchmarked by using them for training on the LUAD – LYM binary dataset. The right side of Figure 1 shows training and validation results obtained by training the BagNet model with no weight initialization. The model is unable to escape from a local optimum during 100 epochs of optimization, the training values are slightly oscillating, while the validation values stay constant. This phenomenon might be caused by unlucky weight initialization, however, three separate runs of this experiment showed similar behavior.

The results of training the BagNet model with the two different pre-trained sets of weights are shown in Figure 2. Both approaches show similar behavior in terms of accuracy and loss, both converging after approximately 60 epochs, however, ImageNet weight initialization leads to slightly better classification performance. This demonstrates the generalizability of deep networks such as ResNet-50, which was able to successfully transfer knowledge learned on the ImageNet dataset to the completely different task of WSI binary classification. The underperformance of LUAD – NEC weights might be caused by the significantly lower training time of this model (approximately 400 epochs) compared to the original BagNet model, which was trained for over 5000 epochs on the ImageNet dataset [5].

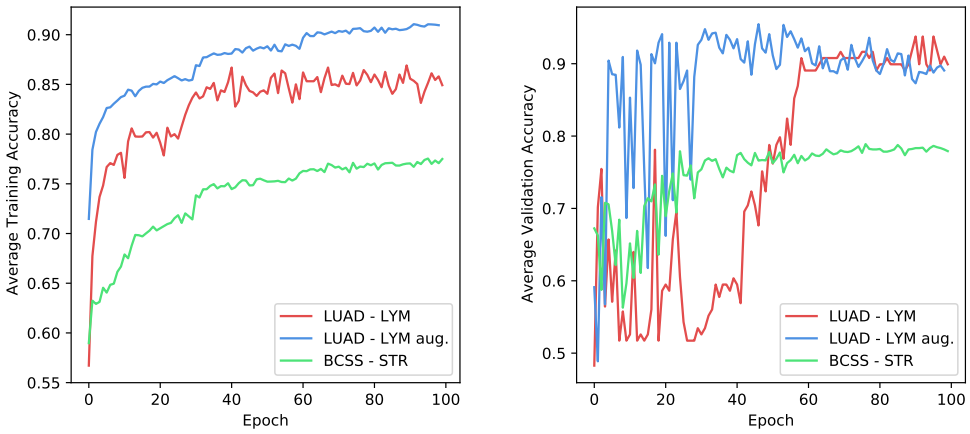

Figure 3: Results of training BagNet with ResNet-50 backbone, ImageNet weight initialization on datasets of different sizes: LUAD – LYM with 1400 training samples and 100 validation samples, augmented LUAD – LYM with 14 000 training samples, and 1000 validation samples, BCSS – STR with 16 000 training samples, and 1600 validation samples.

3.1.4 Dataset properties

The training speed and convergence rate, as well as the ability of neural networks to learn generalization of task-related information, are all dependent on the quality and size of the dataset used for training them. As the eight binary datasets created for the task of binary TME tissue classification are all different in size, we evaluate the effect of dataset size on the training and validation process of the adapted BagNet model. This is done by comparing training and validation accuracy values of three different binary datasets over 100 training epochs.

Figure 3 shows the results of the dataset size examination carried out on LUAD – LYM, augmented LUAD – LYM and BCSS – STR, with 1400, 14 000 and 16 000 training samples respectively. By analyzing the results obtained from this experiment, we managed to draw two different conclusions, one regarding the size of two completely different datasets and one regarding the size difference obtained by using data augmentation.

On the one hand, by comparing the training process on the smallest binary dataset, LUAD – LYM, and one of the largest non-augmented datasets, BCSS – STR, we can observe that a larger dataset leads to a slower and more gradual convergence both in case of training and validation. The smaller dataset ap-

pears to show significantly better validation performance, however, this might be heavily influenced by the significantly smaller size of the validation dataset, 100 samples versus 1600 samples for BCSS – STR. Both validation and accuracy metrics could also be skewed by the fact that the original BCSS dataset was significantly less detailed since it was adapted from a more general dataset [11].

On the other hand, the effects of data augmentation on the training process could also be evaluated by comparing the results on the non-augmented and augmented versions of the LUAD – LYM dataset. Augmenting this small-sized dataset leads to faster convergence and better classification performance including unseen data. However, the model appears to be overfitting the training data after approximately 50 epochs, where the validation accuracy starts showing a descending trend, while the training accuracy is still increasing. This is most likely caused by the low diversity of the augmented dataset, which is a result of applying morphological transformations to the LUAD – LYM binary dataset.

3.2 Quantitative evaluation

In order to benchmark the adapted BagNet architecture for the task of binary classification of TME tissues, the model was first trained on datasets of different sizes with and without using data augmentation, followed by an evaluation on previously unseen data. The results of this experiment are shown in Table 4, along with the sizes of datasets the models were trained and validated on.

By analyzing the performance measure scores obtained in this experiment, we can draw some conclusions about the usability of the BagNet architecture for this specific binary classification task. All models trained on the binary datasets derived from the original LUAD dataset showed promising results with accuracy values starting at 0.94 on previously unseen data, however, the relatively small size of the validation datasets has to be taken into account in this case as well. In the case of the BCSS – STR dataset, which contains a slightly larger validation subset, the BagNet model obtained significantly lower scores than the models trained on the LUAD datasets. This might be caused by the BCSS dataset being less detailed, a topic that is further discussed in Section 3.3.

By comparing the results obtained on augmented datasets with their non-augmented counterparts in the first four rows of Table 4, an increase in accuracy values can be observed in both cases, which indicates that the usage of data augmentation leads to improvements in classification performance.

Dataset	Acc.	Prec.	Recall	Spec.	Train s.	Val. s.
LUAD –						
LYM	0.94	0.925	0.925	0.95	1400	100
LYM aug.	0.954	0.966	0.9175	0.9783	14 000	1000
NEC	0.975	0.9091	1.0	0.9667	6000	40
NEC aug.	0.985	0.9434	1.0	0.98	60 000	400
TE aug.	0.9467	1.0	0.9135	1.0	80 000	600
BCSS –						
STR	0.7888	0.8368	0.7732	0.8085	16 000	1600

Table 4: Results of training BagNet with ResNet-50 backbone with ImageNet weight initialization on binary datasets of different sizes with and without data augmentation. The values shown in this table are the average validation accuracy, precision, recall, and specificity, along with training and validation subset sizes for different binary datasets.

Moreover, the results presented in this table demonstrate comparable performances to radiology experts, in most cases even exceeding the capabilities of human pathologists [14], which shows real-world applicability. However, the robustness of the model on WSI-specific artifacts, color variations, or slide deteriorations mentioned in Section 1 has not been tested, which might limit model usability in real-world environments.

The BagNet model has also been compared to the non-interpretable convolutional networks VGG19 and the backbone used for all BagNet models in this paper, ResNet50. We trained and validated these models on the LUAD – NEC binary dataset with and without transfer learning, and summarized the obtained results in Table 5 using four different performance metrics. Although the added interpretability of the BagNet architecture often leads to slightly worse results compared to its backbones [5], there are cases where interpretable models outperform their non-interpretable counterparts [6, 23]. This phenomenon can also be observed in this experiment, where the BagNet–17 model shows the highest values across all performance measures. This could either be caused by the simplicity of the binary classification problem, or the different nature of tissue scans compared to natural images from ImagNet.

3.3 Qualitative evaluation and interpretability

As our approach is intended to replace human specialists in the safety-critical healthcare industry, one of the main reasons for choosing the BagNet archi-

Model	Acc.	Prec.	Recall	Spec.
Random weights				
VGG19	0.8	0.625	0.5	0.9
ResNet50	0.925	0.8182	0.9	**0.9333**
BagNet–17	**0.95**	**0.8333**	**1.0**	**0.9333**
ImageNet weights				
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**

Table 5: Results of training non-interpretable CNNs for the binary classification task. The values shown in this table are the average validation accuracy, precision, recall, and specificity scores obtained after training the models on the LUAD – NEC dataset with and without using data augmentation.

tecture was its inherent explainability. To demonstrate expected behavior, we hereby evaluate the decision-making process of our models, which in the case of the adapted BagNet architecture consists in the analysis of the models' pixel activation heatmaps. As all of our datasets contain expert annotated segmentation masks for both validation and testing, the exact location of tissue types on WSIs can be leveraged to examine the validity of heatmap pixel activations, thus making it possible to quantitatively measure model interpretability and accordance with expected behavior.

In this experiment, we evaluated trained models on specific WSI patches from the test subsets of different binary datasets and measured in what amount the decisions were based on expected premises. This was done by identifying the top 5% most important heatmap patches, i.e. the ones which contribute most in inferring class evidence, and calculating what percentage of these were located inside the corresponding segmentation mask. Representative examples for this evaluation are shown in Figures 4 and 5. Both of these figures illustrate model performance using four images divided clockwise as follows: the model input WSI patch is visible on the upper left, the segmentation mask for the specific class is shown in the upper right image, the lower right image shows the pixel activation heatmaps for the trained model with a pale outline of the segmentation mask, and the image on the lower left illustrates the top 5% most influential pixels colored green if they lie inside the segmentation mask and yellow otherwise. The most important heatmap patches are illustrated in red.

Specific examples from datasets of different sizes are shown in Figure 4. The difference in WSI qualities between the LUAD and BCSS datasets is visible when comparing segmentation masks of the BCSS – STR example with the other examples from the LUAD datasets. The magnification of the BCSS samples appears to be significantly greater, thus resulting in less detailed samples and segmentation masks as well. This observation should be taken into account when evaluating BagNets' performance on this dataset, which although showing 67.26% of the most important patches in their expected location, appear to be heavily concentrated in the edge of the sample.

When evaluating the other three examples from this figure, the BagNet models appear to be showing promising behavior. The examples from the LUAD – LYM and LUAD – TE datasets demonstrate a decision-making process that aligns with the expected segmentation masks with 80.38% and respectively 89% of the most important patches being inside of their expected locations. The LUAD – NEC example shows slightly weaker performance in this context, however, this might be caused by this class being inherently more difficult to identify than others. Therefore, because of the possible difference in classification difficulty for different datasets, conclusions can not be drawn about the influence of dataset size on the decision-making process of the models.

Model generalizability can be improved by augmenting training data [15]. Therefore, we examine the effects of data augmentation on the decision-making process of the BagNet models by analyzing Figure 5, where examples for two different classes are shown with and without data augmentation. By comparing the two heatmaps of any one of the two classes side-by-side, the confidence-increasing effect of using data augmentation becomes apparent. The patches on the examples without data augmentation are significantly more blurred than on the ones with augmentation, especially in the case of the LUAD – NEC example. This effect might be a consequence of the low diversity dataset expansion resulting from augmentation. Thus, the augmented dataset contains multiple slightly different versions of all images, which could inherently increase classification confidence.

In terms of the more quantitative metric of the percentage of most important patches in their expected location the effects of data augmentation are not as clear. In the case of the LUAD – NEC example, which showed the worst performance in Figure 4, data augmentation leads to significant improvements in terms of decision making with an increase from 39% to 62.64% of most important patches in their expected location. However, applying data augmentation on the smaller LUAD – LYM dataset leads to a slightly worse value for this metric, with an effect of moderately spreading the most important

patches over the image. Apart from this minor decay in expected behavior, we overall advise the usage of data augmentation for the task of TME tissue binary classification because of the effects of increased confidence and accuracy improvements shown in Section 3.2.

4 Conclusion and discussion

The automation of tasks performed by human pathologists poses various challenges. For instance, a machine learning model trained on data from one hospital may perform poorly on data from another hospital due to differences in scanning equipment and tissue processing protocols. Medical datasets also require expert annotation, which is time-consuming and resource-intensive, leading to a scarcity of high-quality annotated data. Moreover, the black-box nature of deep learning models is also a significant challenge in medical image processing, where the interpretability and explainability of algorithms are becoming essential.

The focus of this paper lies in solving the cancer detection problem of identifying differences in tissue types of tumor microenvironments. Our approach to TME tissue type detection involves breaking it down into multiple binary classification problems. This method can be used in conjunction with or as a replacement for human radiologists to label whole-slide images, which can later be used as inputs for more complex segmentation techniques.

To accomplish binary tissue classification, we introduced a modified version of the explainable BagNet architecture. The proposed model is capable of outperforming human experts in different binary classification tasks, providing a viable alternative to human-based tissue detection. Using the BagNet network architecture also addresses concerns surrounding the back-box nature of medical image processing models by being able to justify its decisions. In addition, we also demonstrated that the decision-making process of these models is also aligned with that of human radiologists. Furthermore, we addressed the challenge of limited medical data by augmenting our binary datasets, and we analyzed the effects of augmentation techniques on model performance and decision-making validity.

Being able to accurately and reliably identify TME tissue types without the need for the involvement of human experts presents a variety of different advantages in the field of medical imaging. Firstly, it leads to the burden of disease diagnosis being taken off the shoulders of pathologists, them being able to focus on patient treatment and care. Secondly, this could also lead to new

developments in more complex classification and segmentation tasks regarding tumor microenvironments. For instance, our trained models could also be used for creating large datasets annotated with patch-level classification labels of tissue types, thus providing future models with sufficiently large annotated datasets without the necessity for human involvement. Moreover, these models could provide a way to fully automate more complex approaches that rely on human labeling, such as the method introduced in [11].

The first future research direction possibly worth exploring regarding the proposed models is the evaluation of the benefits these create in terms of new developments in this field. This could be done by using the models to create annotated datasets, which would then be used for training and benchmarking various state-of-the-art approaches in TME image processing. The usability of these models as initial steps for more complex methods should also be evaluated by carrying out experiments comparing performances with and without using them. Another related question concerns the robustness of the models for WSI differences caused by artifacts, discoloration, and tissue deterioration. This characteristic could be reviewed by evaluating the models on new datasets obtained at different laboratories with slight variations in tissue sample acquisition pipelines.

Further research regarding this work could be carried out in the direction of model decision analysis. Although heatmaps can be reviewed qualitatively, there is a requirement for a more objective quantitative evaluation. The process of location of the most relevant patches carried out in this paper represented one quantitative evaluation method, however, more complex and representative quantities could also be analyzed, such as the relevance mass accuracy and relevance rank accuracy metrics proposed in [3]. Moreover, model performance could be evaluated on larger validation and testing datasets as well to further solidify results regarding classification accuracy.

Acknowledgements

This work was supported by Babeş–Bolyai University's Special Scholarship for Scientific Activity (No. 36586/25.11.2022) and the Collegium Talentum Programme of Hungary. A part of the research leading to these results was supported by the Márton Áron College of ELTE Eötvös Loránd University. We thank Dr. Bognár Gergő and Dr. Kovács Péter from the Faculty of Numerical Analysis, ELTE Eötvös Loránd University for providing the computer hardware used in the experiments.

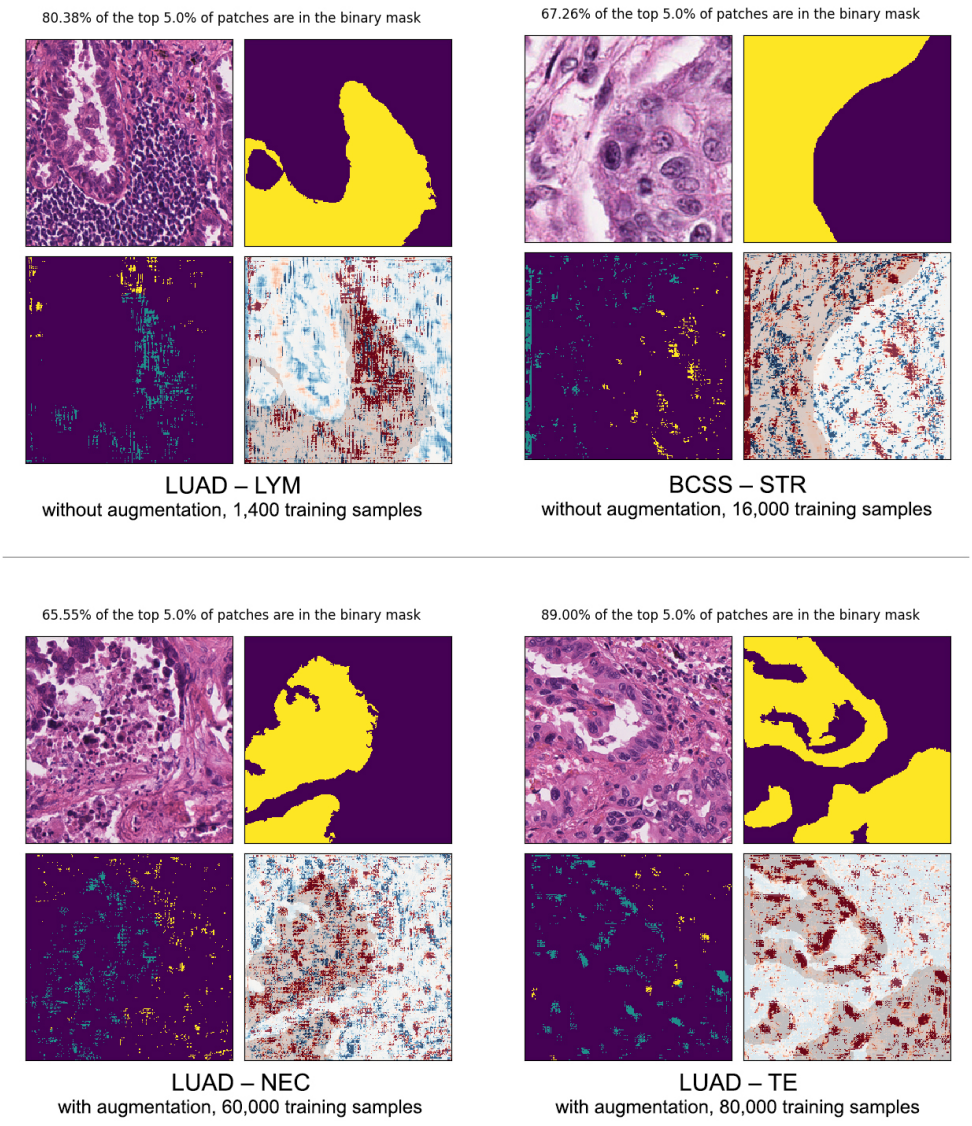

Figure 4: Heatmap analysis of BagNet models trained on four different datasets with varying sizes.

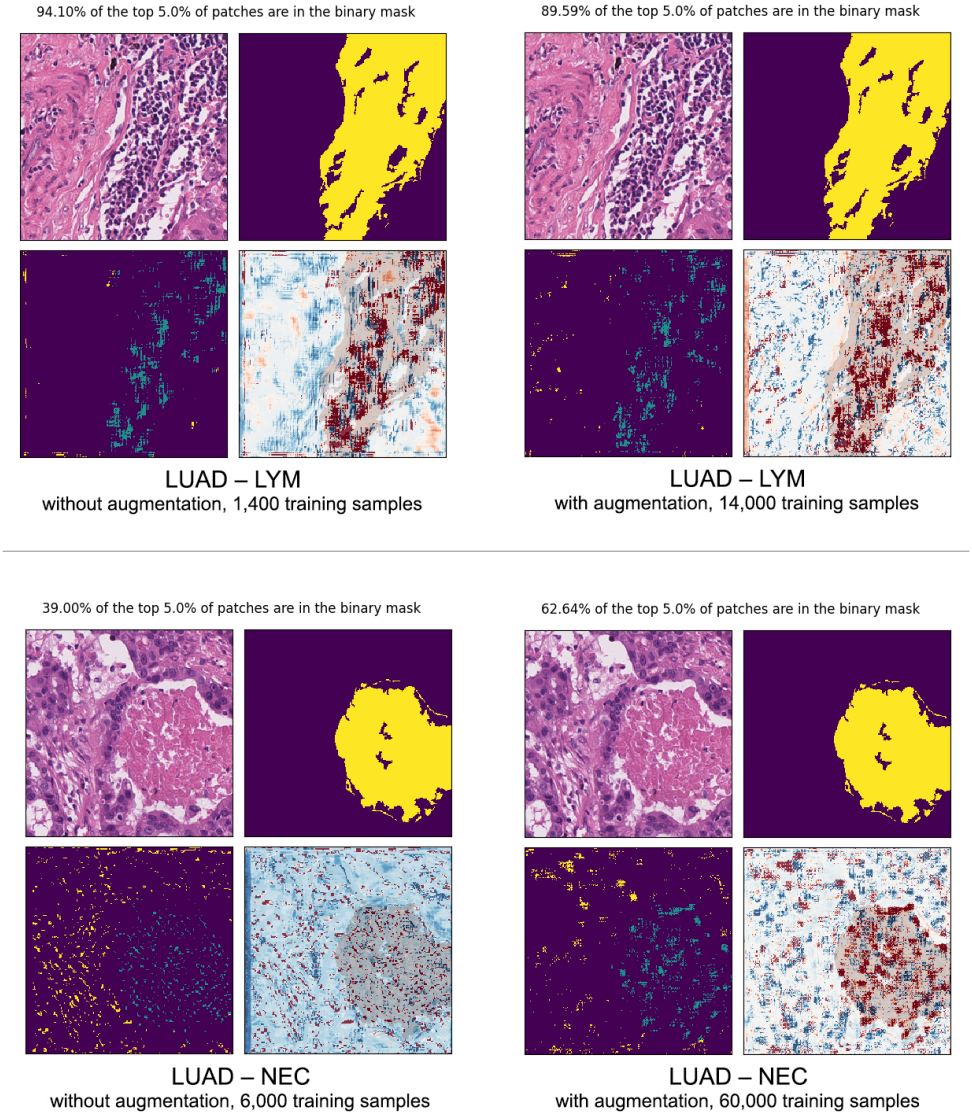

Figure 5: Heatmap analysis of BagNet models trained for binary classification of two different classes with and without data augmentation.

References

[1] K. AbdulJabbar, S. E. A. Raza, R. Rosenthal, M. Jamal-Hanjani, S. Veeriah, A. Akarca, T. Lund, D. A. Moore, R. Salgado, M. Al Bakir, L. Zapata, Geospatial immune variability illuminates differential evolution of lung adenocarcinoma, *Nature Medicine* 26(7), 2020, pp. 1054–1062. ⇒61

[2] M. Amgad, H. Elfandy, H. Hussein, L. A. Atteya, M. A. Elsebaie, L. S. Abo Elnasr, R. A. Sakr, H. S. Salem, A. F. Ismail, A. M. Saad, J. Ahmed, Structured crowdsourcing enables convolutional segmentation of histology images, *Bioinformatics* 35(18), 2019, pp. 3461–3467. ⇒61, 63, 64

[3] L. Arras, A. Osman, W. Samek, CLEVR-XAI: a benchmark dataset for the ground truth evaluation of neural network explanations, *Information Fusion* 81, 2022, pp. 14–40. ⇒76

[4] M. S. Ayhan, L. B. Kümmerle, L. Kühlewein, W. Inhoffen, G. Aliyeva, F. Ziemssen, P. Berens, Clinical validation of saliency maps for understanding deep neural networks in ophthalmology, *Medical Image Analysis* 77, 2022, p. 102364. ⇒68

[5] W. Brendel, M. Bethge, Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet, *arXiv preprint* arXiv:1904.00760, 2019. ⇒62, 63, 66, 67, 68, 69, 72

[6] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, J. K. Su, This looks like that: deep learning for interpretable image recognition, *Advances in neural information processing systems* 32, 2019. ⇒72

[7] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248-255). ⇒63

[8] I. B. Dimenstein, Grossing biopsies: an introduction to general principles and techniques, *Annals of Diagnostic Pathology* 13(2), 2009, pp. 106–113. ⇒61

[9] K. Doi, Current status and future potential of computer-aided diagnosis in medical imaging, *The British Journal of Radiology* 78, 2005, pp. 3–19. ⇒61

[10] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, T. Chen, Recent advances in convolutional neural networks, *Pattern Recognition* 77, 2018, pp. 354–377. ⇒67

[11] C. Han, J. Lin, J. Mai, Y. Wang, Q. Zhang, B. Zhao, X. Chen, X. Pan, Z. Shi, Z. Xu, S. Yao, Multi-layer pseudo-supervision for histopathology tissue semantic segmentation using patch-level classification labels, *Medical Image Analysis* 80, 2022, p. 102487. ⇒61, 62, 63, 71, 76

[12] D. Hanahan, R. A. Weinberg, Hallmarks of cancer: the next generation, *Cell* 144(5), 2011, pp. 646–674. ⇒61

[13] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778. ⇒67

[14] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, J. H. Saltz, Patch-based convolutional neural network for whole slide tissue image classification, *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2424–2433. ⇒72

[15] N. Kanwal, F. Pérez-Bueno, A. Schmidt, K. Engan, R. Molina, The devil is in the details: whole slide image acquisition and processing for artifacts detection, color variation, and data augmentation: a review, *IEEE Access* 10, 2022, pp. 58821–58844. ⇒61, 74

[16] S. Morales, K. Engan, V. Naranjo, Artificial intelligence in computational pathology -– challenges and future directions, *Digital Signal Processing* 119, 2021, p. 103196. ⇒61

[17] C. Park, H. I. Suk, Deep joint learning of pathological region localization and Alzheimer's disease diagnosis, *arXiv preprint* arXiv:2108.04555, 2021. ⇒63, 68

[18] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, ImageNet large scale visual recognition challenge, *International Journal of Computer Vision* 115, 2015, pp. 211–252. ⇒63

[19] J. Su, D. V. Vargas, K. Sakurai, One pixel attack for fooling deep neural networks, *IEEE Transactions on Evolutionary Computation* 23(5), 2019, pp. 828–841. ⇒61

[20] A. Taleb, W. Loetzsch, N. Danz, J. Severin, T. Gaertner, B. Bergner, C. Lippert, 3D self-supervised methods for medical imaging, *Advances in Neural Information Processing Systems* 33, 2020, pp. 18158–18172. ⇒61, 63

[21] E. Tjoa, C. Guan, A survey on explainable artificial intelligence (XAI): Toward medical XAI, *IEEE Transactions on Neural Networks and Learning Systems* 32(11), 2020, pp. 4793–4813. ⇒61

[22] B. H. Van der Velden, H. J. Kuijf, K. G. Gilhuijs, M. A. Viergever, Explainable artificial intelligence (XAI) in deep learning-based medical image analysis, *Medical Image Analysis*, 2022, p. 102470. ⇒61

[23] J. Wang, H. Liu, X. Wang, L. Jing, Interpretable image recognition by constructing transparent embedding space, *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 895-904. ⇒72

[24] M. Wang, D. Zhang, D. Shen, M. Liu, Multi-task exclusive relationship learning for Alzheimer's disease progression prediction with longitudinal data, *Medical Image Analysis* 53, 2019, pp. 111–122 ⇒61

[25] C. Wang, X. Zhu, J. C. Hong, D. Zheng, Artificial intelligence in radiotherapy treatment planning: present and future, *Technology in Cancer Research & Treatment* 18, 2019, p. 1533033819873922. ⇒61

Received: May 9, 2023 • Revised: June 19, 2023