

TDK-dolgozat

Galiger Gergő

Babeş–Bolyai Tudományegyetem
Matematika és Informatika Kar

Histopathological tissue analysis using explainable deep neural networks and data augmentation techniques

Szerző:

Galiger Gergő

Babeş–Bolyai Tudományegyetem, Matematika és Informatika Kar, Adatelemzés és modellezés
szak, II. év

Témavezető:

dr. Bodó Zsolt, egyetemi docens

Babeş–Bolyai Tudományegyetem, Matematika és Informatika Kar

Babeş–Bolyai Tudományegyetem

Matematika és Informatika Kar

Hisztopatológiai szövetvizsgálat interpretálható mély neurális hálók és adataugmentációs módszerek segítségével

Szerző:

Galiger Gergő

Babeş–Bolyai Tudományegyetem, Matematika és Informatika Kar, Adatelemzés és modellezés
szak, II. év

Témavezető:

dr. Bodó Zsolt, egyetemi docens

Babeş–Bolyai Tudományegyetem, Matematika és Informatika Kar

Contents

1	Introduction	1
1.1	Medical image processing	1
1.2	Whole-slide imaging	2
1.3	Challenges of computer-aided diagnosis	4
1.4	Explainable tissue analysis with data augmentation	5
2	Theoretical Background	7
2.1	Model architecture	7
2.1.1	Convolutional neural networks	7
2.1.2	Explainable machine learning	8
2.1.3	Bag-of-local-Features – BagNet	9
2.2	Data augmentation	11
3	Methodology	13
3.1	Background motivation	13
3.2	Weakly-annotated datasets	13
3.3	Problem formulation	15
3.4	Dataset expansion	16
3.5	Implementation details	17
4	Experiments and Results	18
4.1	Training process	18
4.1.1	Model backbone	18
4.1.2	Transfer learning	19
4.1.3	Dataset properties	20
4.2	Quantitative evaluation	22
4.3	Qualitative evaluation	23
5	Conclusion	27
5.1	Overview	27
5.2	Theses	27
5.3	Discussion	28

1 Introduction

1.1 Medical image processing

With the continuous digitization of all scientific fields, the analysis of the collected data plays an increasingly important role in solving problems that were previously thought to only be solvable by humans. Medical image processing is a rapidly evolving field that plays a critical role in modern medicine. With the increasing use of imaging technologies such as X-rays, CT scans, MRIs, ultrasound, and whole-slide images, there is a growing need for specialized algorithms to analyze these images accurately and efficiently [50]. The accurate interpretation and analysis of medical images can aid in the diagnosis of various diseases and conditions [10], provide information about the progression of diseases [42, 60], and assist in treatment planning [24, 59].

Because of its sheer complexity, distinct anomalies at different magnification levels of a living organism require different imaging methods. For example, there is an apparent difference between capturing large, organ-level abnormalities, like kidney stones, and cellular-level anomalies, like cancerous cells in the lung. Therefore, the field of medical imaging encompasses a wide range of specific techniques that are adequate for different types of diseases and diagnoses [48]. The amount of detail required for a correct analysis is another factor that further differentiates the medical data acquisition methods [28].

The majority of the most widely-used medical imaging methods fall into the category of non-invasive techniques [23]. These approaches do not require removing tissue samples from patients and are mostly preferred due to their safety and convenience [13]. A widely known example in this category is X-ray radiography, which uses electromagnetic radiation passing through the body to obtain an image based on the absorption and attenuation levels of various density tissues [61]. While being widely used by itself in medical image processing, this method also provides the basis of X-ray computed tomography (CT), which combines X-ray imaging with computer software to produce images of cross sections of the human body [21]. Other non-invasive imaging techniques include magnetic resonance imaging (MRI) and ultrasound imaging, which rely on magnetic and radio frequency fields, and high-frequency sound waves respectively for producing medical images [23].

Different methods for processing images obtained from non-invasive techniques have been successfully used in various real-world applications. Images obtained with X-ray radiography

were used for bone fracture detection on field programmable gate array (FPGA) hardware in [35]. The authors of [33] proposed a machine learning algorithm to diagnose coronavirus disease (COVID-19) by identifying abnormal patterns of CT and X-ray images. Computed tomography images are also adequate for coronary artery disease and flow assessment [27, 55]. MRI images are frequently analyzed for diagnosis of autism spectrum disorder [38], schizophrenia [12] and different mood disorders [25], but different types of cancer as well [43, 8]. Medical ultrasound has also been used in a variety of different fields of pathology, from measuring spine curvature [63], muscle disease detection [62], to all kinds of different medical image processing applications [19].

In spite of the undoubtable usefulness of non-invasive medical imaging methods, all of the previously listed approaches have their own limitations [46]. One of the major problems with non-invasive imaging is the limited accessibility of these techniques. Parts of the body with difficult access, such as deep-seated tumors or structures behind bones, are hard to access with non-invasive methods. Another issue with these approaches arises when there is a need for high precision and resolution of the examined body part. For example, high cellular-level resolution is, with a few exceptions [44], not obtainable by only using non-invasive techniques. For applications involving high-resolution capturing of deep-seated parts of the human body, such as precise cancer cell detection and analysis tasks, performing a biopsy, i.e. removing tissue samples from the targeted body parts is essential.

1.2 Whole-slide imaging

Although not as convenient as non-invasive medical imaging methods, invasive techniques offer clear advantages in terms of accessibility of deep parts of the body and cellular-level resolution. This is achieved by removing specific tissue samples of the body through some form of biopsy. Needle biopsy is the most commonly used method for taking samples from inside the body, but hard-to-reach areas require more complex techniques, such as endoscopy or cystoscopy [9]. When performed within professional guidelines, such methods are also considered safe, with high success rates and low rates of complication [26].

Invasive medical imaging approaches involve a series of complex procedures that require specialized expertise and sophisticated tools. Obtaining the tissue sample from the patient's body through biopsy represents only the initial step of a professional image acquisition pipeline. As shown in Figure 1, the collected samples go through a series of tissue processing steps from

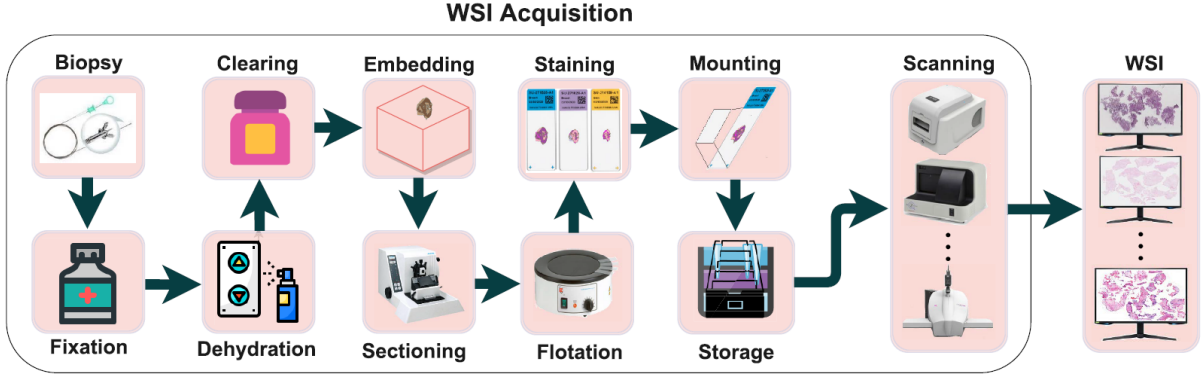


Figure 1: The WSI acquisition pipeline is a complex procedure. In order to obtain quality images, a variety of different steps are necessary, which are all prone to material and human errors that degrade tissue sample quality [22].

fixation through dehydration and cleaning to staining to obtain high-quality digitized images [22]. Each step of this pipeline ensures that the processed tissue sample is adequately prepared for analysis and retains its properties through storage. Finally, for a more in-depth analysis, prepared samples are digitized using adequate scanning equipment, producing digital images called whole-slide images (WSI) [52].

The scanned WSIs represent high-resolution images of gigapixel order with physical sizes often exceeding tens of gigabytes. This makes it even possible to emulate microscope navigation and magnification on a digitized tissue sample, as illustrated in Figure 2. Because of their large space requirements, the computational cost of analyzing single WSIs with different algorithms is very high, and so are the memory requirements of a system capable of carrying out such an analysis [34]. Furthermore, files of this size do not even fit on a GPU, which renders the task of using machine learning methods for diagnosis almost impossible. 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 [22].

The process of digitization of tissue samples opened new possibilities for automating different tasks in the field of medical image processing [22], especially digital histopathology, which refers to the examination and diagnosis of cancer and other cellular-level diseases. The introduction of WSIs made it possible to develop different computer-aided diagnosis (CAD) and prognosis systems for this purpose. The most common image processing tasks with CAD include classification, segmentation, and detection [30].

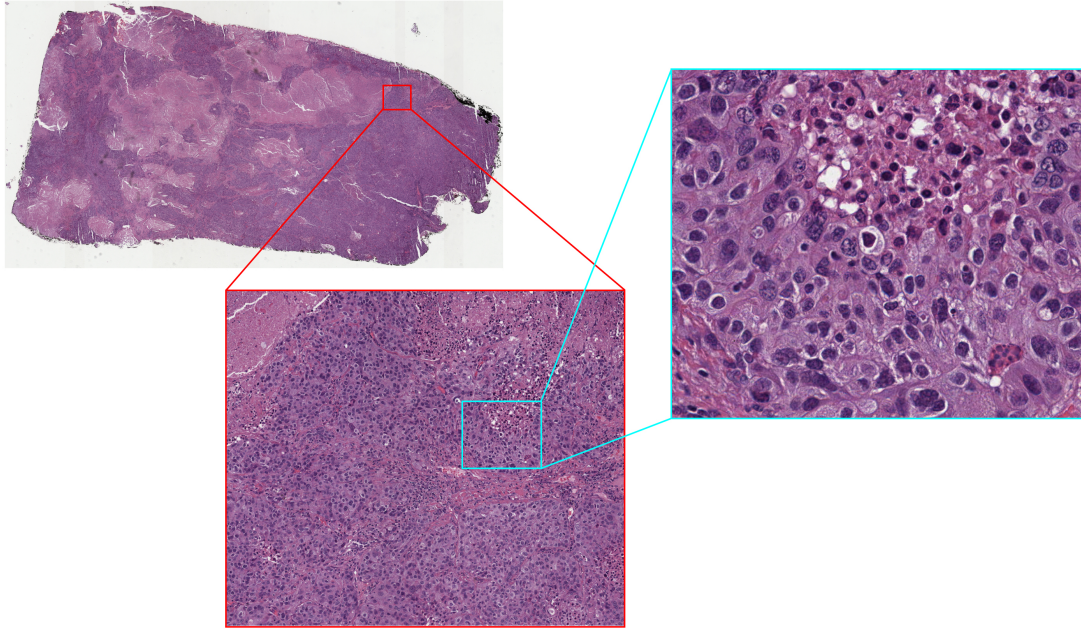


Figure 2: A whole slide image of lung tissue from the TCGA open access dataset [57]. These images contain a high level of detail even with extreme magnification due to their high resolution and large size.

1.3 Challenges of computer-aided diagnosis

Automating different tasks of trained human pathologists using CAD systems poses a variety of different challenges by itself. For example, a machine learning model for WSI processing tasks trained on data collected from one hospital is likely to have comparable results to human radiologists on a test set collected from the same hospital. But when this model is evaluated on data from a different hospital, with slightly different scanning equipment and tissue processing protocols, the subtle differences in the obtained images cause the performance of the model to degrade significantly [41]. Machine learning-based methods are even vulnerable to only changing a single pixel of an image [51]. In the field of WSI processing anomalies significantly larger than one-pixel differences are common because of the complexity of the WSI acquisition pipeline. Different artifacts (air bubbles, dirt) may contaminate tissue samples, color variations are caused by different staining substances, and deterioration (tears, cracks) of the samples is also common [22].

One of the other main obstacles that impede progress in this field, is the difficulty in creating a large, high-quality annotated medical dataset, which is essential for all tasks involving a type of supervised learning. Unlike natural scene image data, medical datasets require experts with domain knowledge to perform the annotation process, making it a highly specialized and resource-intensive task. Correct annotation of these high-resolution images 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 [54]. These factors collectively lead to a scarcity of annotated medical data, both in terms of volume and annotation quality. However, different approaches are constantly developed and evaluated for solving the problem of medical data scarcity, such as semi-supervised [7] and weakly-supervised learning [15] and different data augmentation techniques [11].

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 [58]. In the context of medical decision-making, the stakes are high as human lives are at risk. Although deep learning methods have shown promising performances in medical tasks, their black-box nature has raised concern among medical experts [56]. Furthermore, the EU GDPR regulations also require that patients have the right to meaningful information about how a decision regarding their diagnosis was made [58]. Thus, the interpretability and explainability of algorithms used in medical image processing have become pressing issues. This led to the increasing usage and evaluation of interpretable deep learning or explainable artificial intelligence (XAI) approaches, which involve providing an explanation for the decision-making process of these models [36].

1.4 Explainable tissue analysis with data augmentation

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 [16] and also affects the prognosis and response to therapy of cancer patients [1]. We reformulate the task of TME tissue type detection as multiple binary classification problems, that can be used complementarily or even as a replacement for human radiologists in labeling initial data required for more complex segmentation approaches, such as the one proposed in [15].

For solving the task of binary tissue classification, we propose an adapted version of the explainable BagNet architecture, which demonstrates matching and in some cases even exceeding 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 by using data

augmentation techniques, the effects of which on both the performance and validity of model decisions are further analyzed in detail.

The rest of this paper is structured as follows: Section 2 summarizes the theoretical background regarding the BagNet model architecture and data augmentation techniques. Section 3 describes the specific TME binary classification task formulation and necessary adaptations in both model architecture and dataset structure. Specific datasets information and further implementation details are also listed here. In Section 4, 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 5, where we also provide further research directions regarding this topic.

2 Theoretical Background

2.1 Model architecture

2.1.1 Convolutional neural networks

Inspired by the natural visual perception mechanisms, convolutional neural networks (CNN) represent a widely utilized deep learning architecture [14]. The modern framework of CNNs was introduced as a multi-layer artificial neural network called LeNet-5, which could classify hand-written digits [29]. Although there are many different nuances to CNN architectures, their basic components are very similar. These components were also present in the initial LeNet-5 architecture illustrated in Figure 3.

Three main types of layers consist the basis of the majority of convolutional neural networks: convolutional, pooling, and fully-connected layers [14]. The convolutional layers are composed of multiple convolution kernels which are used to compute different feature maps in order to learn feature representations of inputs. New feature maps are obtained by applying an element-wise activation function to convolution results, such as sigmoid or ReLU. The pooling layers reduce the resolution of feature maps with the main objective of achieving shift invariance. This is commonly done by using average pooling or max pooling.

Convolutional and pooling layers are the main components of convolutional neural networks [17]. More general CNNs commonly include fully-connected layers in their architecture to perform high-level reasoning. A fully-connected layer connects all neurons of two layers to obtain global semantic information, however, these can be replaced by 1×1 convolution operators [31]. The final layer of all CNNs is an output layer that maps previous layer activations to the task space at hand. Commonly used methods in output layers include sigmoid and softmax functions, and support vector machines (SVM). As with all machine learning algorithms, training a CNN is a global optimization problem with the aim of finding the best network parameters by minimizing a loss function. Stochastic gradient descent is commonly used for the task of optimizing convolutional neural networks [14].

Since the first introduction of the LeNet-5 architecture, many different methods have been developed to overcome obstacles regarding CNN performance and training. Notable improvements have been made by many CNN architectures, such as AlexNet [45], VGGNet [49], or GoogleNet [53], which proposed deeper and deeper CNN structures. Increasing the depth of CNNs results in a better ability to learn feature representations, however, it also leads

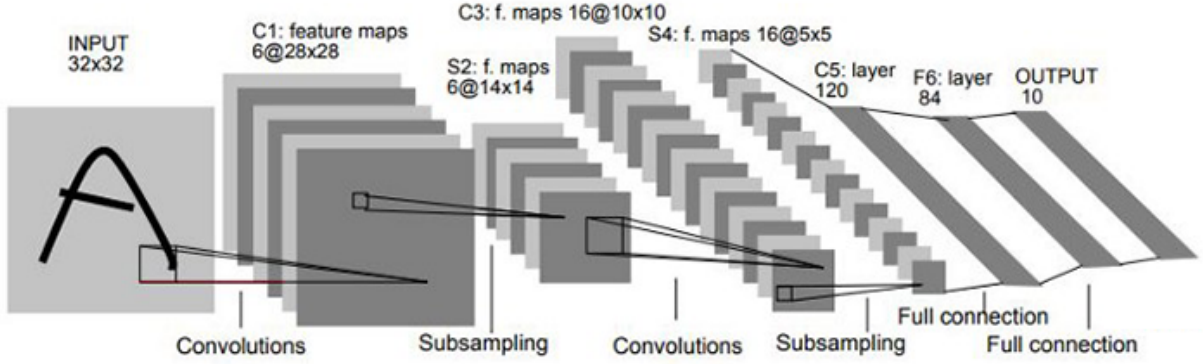


Figure 3: The architecture of the famous LeNet-5 CNN [29].

to optimization difficulties due to increased network complexity [14]. With the introduction of skip-connections in the ResNet residual CNN architecture, the authors were able to successfully optimize very deep networks and obtain great results on ImageNet classification tasks [17]. ResNet and other CNN architectures are commonly used as backbones for more specialized models intended for solving specific problems, even in medical image processing: [5, 20, 32].

2.1.2 Explainable machine learning

As discussed earlier, the interpretability and explainability of machine learning models are of critical importance in applications like medical image processing. Explainability in the context of neural networks refers to providing additional information about the model's decision, which can be achieved in a variety of different ways [36]. Simpler deterministic models, such as an if-else rule-based shallow decision tree, are inherently explainable and have a decision-making process that is relatively easy to understand. However, currently prominent models in the field of medical image processing are based on deep neural networks (DNN) with complex layer structures that often provide no information about their decisions [56].

Directly understanding the inner workings of black box DNNs is not possible. However, different techniques have been proposed to understand such models' decision-making processes, such as visualization by perturbation, visualization by gradient/backpropagation, or explanation by example [36]. Likewise, heatmaps are also commonly used for the representation of hidden information in DNN neuron activations by analyzing the importance of certain input features. Some aspects of the inner workings of DNNs can be identified by carrying out visual inspection and perturbation analyses on such heatmaps [3].

One particular approach to image processing tasks that allows the usage of heatmaps for

decision interpretation by design is the bag-of-features (BoF) representation. In analogy to bag-of-words representations, bag-of-features involve counting the number of occurrences of certain features in an image, i.e. visual words. A vocabulary of these visual words is first created, similar to a vocabulary of important words in bag-of-words representations. The counts of each visual word in the vocabulary are assembled as a term vector, which serves as the input to a classifier such as SVM or MLP. The representation of image features as term vectors results in the complete disregard of the spatial relationships between features. This approach has been widely used in successful image classification models [37]. The most effective interpretation of BoF models can be achieved by using a linear classifier on top of the term vector. In this case, the impact of a specific part of the input on the classifier is not influenced by the rest of the input [6].

2.1.3 Bag-of-local-Features – BagNet

In contrast to basic BoF models that use a simple spatial aggregation of image features, or visual words, deep neural networks utilize information across an entire image in a non-linear manner. The BagNet model family, introduced in [6], demonstrates the potential of combining the flexibility and performance of DNNs with the interpretability of BoF models. BagNets are linear DNN-based BoF models with a strong resemblance to the ResNet-50 fully convolutional network architecture. The differences between the two structures only lie in the replacement of some stride values and many 3×3 with 1×1 convolutions, thus limiting the receptive size of the topmost convolutional layer to $q \times q$ pixels.

For effective interpretability, a linear classifier is applied to the outputs of the last residual blocks to infer the class evidence of each patch corresponding to the $q \times q$ receptive fields. This results in a heatmap of patch activations for each class, which is then averaged across all patches and passed through a softmax operator to infer the image-level class evidence (i.e. logits). The resulting architectures were named after their topmost receptive fields as BagNet- q , where $q \in [9, 17, 33]$ were further evaluated. The precise structures of these three architectures are illustrated in Figure 4.

Limiting the topmost receptive fields of the model results in the BagNet architecture adapting the BoF characteristic of neglecting spatial relationships between image parts. This means that mixing around image patches has no effect on the final decision of BagNet, hence the name Bag-of-local-Features. Using this insight, the BagNets have been used to uncover

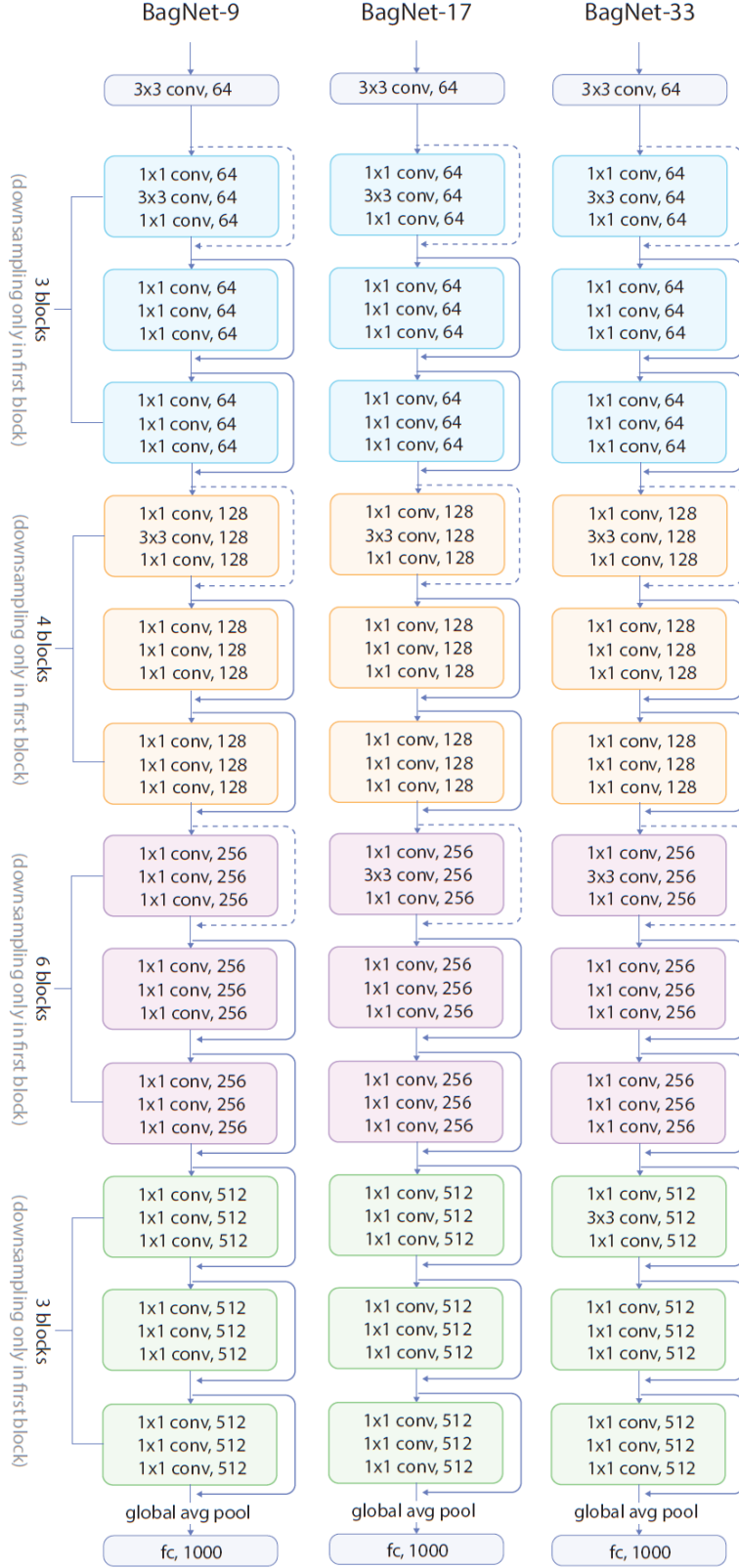


Figure 4: The architecture of the BagNet explainable neural network [6]. As BagNet was inspired by the ResNet-50 residual architecture, their structures are almost identical. The only differences lie in some minor changes in stride values and the replacement of most 3×3 convolutions with 1×1 convolutions.

similarities between the decision-making processes of BoF models and popular DNNs like VGGNet or ResNet [6]. Furthermore, these models have already been successfully used in a few medical image processing tasks [5, 39, 40]. In [20], the authors conclude that BagNets may be a compelling alternative to standard DNNs for medical image analysis tasks since they do not require the implementation of additional explainability methods.

2.2 Data augmentation

The success of applying CNNs in different image processing tasks led to the development of many branches of study focusing on improving different aspects of such neural networks. One of the most difficult challenges among all these different approaches lies in finding ways to improve the generalizability of these models [47]. The generalizability of a certain model is characterized by the difference in performance on training data versus previously unseen data, i.e. test data. A model with poor performance on unseen data is not capable of abstraction beyond training data and thus has poor generalizability. This issue is in practice often caused by insufficient information in the training dataset, which can be a direct consequence of a small dataset size [22]. As mentioned previously, the phenomenon of insufficient data is a common problem in medical imaging.

One of the most straightforward methods of improving model generalizability and reducing overfitting is data augmentation, which in short refers to the enhancement of the size and quality of training datasets. This approach is based on the premise that augmentations can extract additional information from the original dataset. An augmented training dataset represents a broader range of possible data points of the problem space, thus reducing the gap between the training and testing sets [47].

The expansion of datasets in medical image processing is mainly achieved through three different data augmentation approaches: morphological transformation, color augmentation, and generative methods. Color augmentation methods introduce new training samples by manipulating the color distribution of an image while preserving its structure. These methods tend to increase generalizability between differently sourced datasets, thus providing a solution for the aforementioned issue of differences in medical image acquisition methods. Generative augmentation methods introduce new image content instead of modifying existing images. Although these approaches are showing promising results, the literature in the context of histopathological images is scarce [22].

Data augmentation by the morphological transformation of images involves changing the shape, structure, or field of view of images [22]. This includes transformations such as rotations, mirroring, shifts, warps, cropping, and others. Unlike natural scene images like landscapes or portraits, tissue samples contained in WSIs are not limited by a certain orientation. For instance, flipping WSIs upside down, or rotating them by extreme angles does not result in a difference in perception of these images. Therefore, morphological augmentation approaches are by intuition adequate for histopathological image processing applications.

3 Methodology

3.1 Background motivation

In the context of histopathological image processing, segmentation tasks are commonly considered substantially more difficult than detection, or even classification. Identifying different regions of an image, the main objective of segmentation tasks inherently encapsulates a type of detection and classification. Therefore, segmentation methods might rely on additional, a-priori information besides the input images. As a related example, in [15], which was the main starting point of this paper, the proposed state-of-the-art tissue segmentation 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 classification labels used as inputs. The specific task thus becomes the detection and differentiation of various tissue types present in the tumor microenvironments (TME) of cancerous tissues, which can be considered an initial step for the segmentation method proposed in [15]. As this approach intends to replace human experts for this specific task, the reliability of its decision-making process is of utmost importance. Therefore, the main emphasis when choosing a suitable architecture for this purpose was the explainability of decisions, hence the usage of interpretable neural networks. Furthermore, data augmentation methods were also leveraged in order to improve the performance and even the decision-making process of used models on small datasets.

3.2 Weakly-annotated 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 and the publication of such datasets also raises concerns regarding patient privacy. The datasets used in this paper originate from the previously mentioned paper [15], where two different datasets of TMEs were created and published for further research:

LUAD-HistoSeg: This dataset was specifically created for training the segmentation

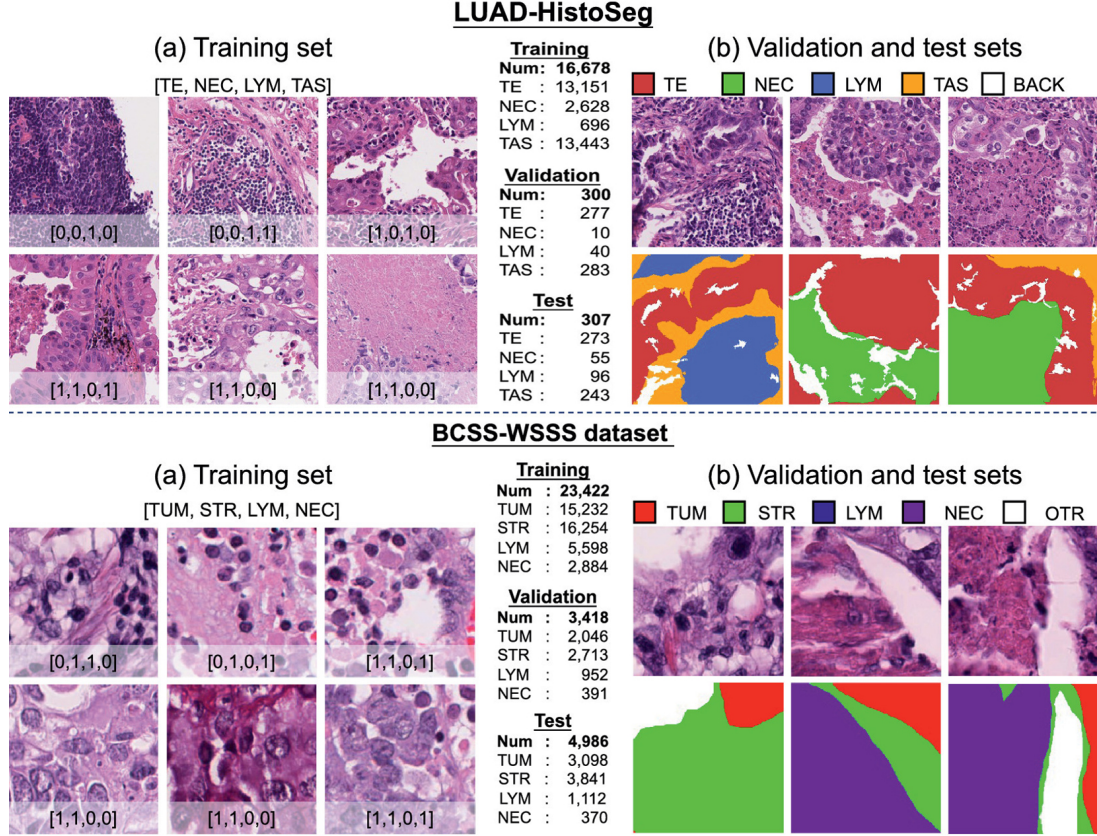


Figure 5: Examples and per-class distribution of the datasets released in [15].

approach described above and contains patches from 54 patients from the Department of Pathology, Guangdong Provincial People’s Hospital suffering from lung cancer. The dataset was built from 29 WSIs for training and 25 WSIs for validation and testing with randomly sampled patches from each with 224×224 resolution and 10X magnification inside the tumor bulk. The final dataset contained a training set of 16 678 patches, a validation set of 300 patches, and a test set of 307 patches with four tissue categories: tumor epithelial (TE), tumor-associated stroma (TAS), necrosis (NEC) and lymphocyte (LYM).

BCSS-WSSS: The second dataset was adapted from a previously published dataset [2] containing patches from 151 different WSIs of breast cancer. The final adapted version of this dataset contained a training set of 23 422 patches, a validation set of 3418 patches, and a test set of 4986 patches including five classes: Tumor (TUM), Stroma (STR), Lymphocytic infiltrate (LYM), Necrosis (NEC) and Other (OTR).

Both datasets published in [15] are weakly annotated, where validation and testing samples were segmented by pathologists (pixel-level annotation), while training samples only contain class labels of whether a certain tissue type is present in the patch or not (patch-level annotation). The per-class distribution of the two datasets and examples of patches and annotations are

shown in Figure 5.

3.3 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, more specifically BagNet-17, which demonstrated the best performance on ImageNet [45] tasks among the originally proposed versions of this architecture [6]. The promising results obtained in a variety of different tasks despite its simplicity, and its inherent explainability with pixel-level activation heatmaps, makes the BagNet architecture a competitive choice for solving this specific problem. As training and validation datasets are fully segmented in the case of both the LUAD and BCSS datasets, 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, 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.

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

Binary dataset	Train set	Test set	Validation set
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.

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.

3.4 Dataset expansion

By examining dataset size and distribution numbers shown in Figure 5, it can be stated that 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 is 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,

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.

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.

3.5 Implementation details

All the convolutional neural networks analyzed in our experiments were implemented in PyTorch and we used the Augmentor¹ Python package for the augmentation of the binary datasets. We trained the models using an NVIDIA Titan Xp GPU and utilized ResNet-50 as the classification backbone for the BagNet-17 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.

¹Augmentor Python Package Documentation

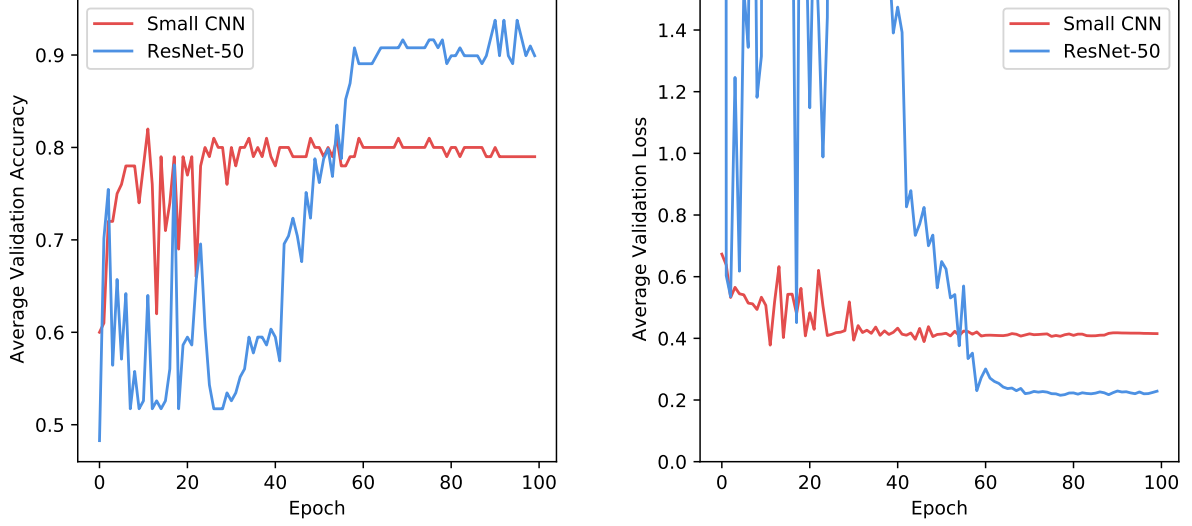


Figure 6: Results of benchmarking small CNN and ResNet-50 backbones for BagNet for binary classification trained on the LUAD – LYM dataset.

4 Experiments and Results

4.1 Training process

4.1.1 Model backbone

As mentioned in Section 2.1.1, neural networks show a tendency to benefit from increasing depth in various image processing tasks. In [6], the BagNet network was proposed with the ResNet-50 architecture as a backbone, which is considered a highly layered architecture by current standards [17]. In order to test the necessity of such a deep backbone architecture for the binary classification task of TMEs, 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 and loss values of the two models are illustrated in Figure 6. Using the small CNN as a backbone, both the accuracy and loss 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 convergence but shows substantially improved classification performance. Therefore, we use the ResNet-50 as the backbone for our BagNet models in the following experiments.

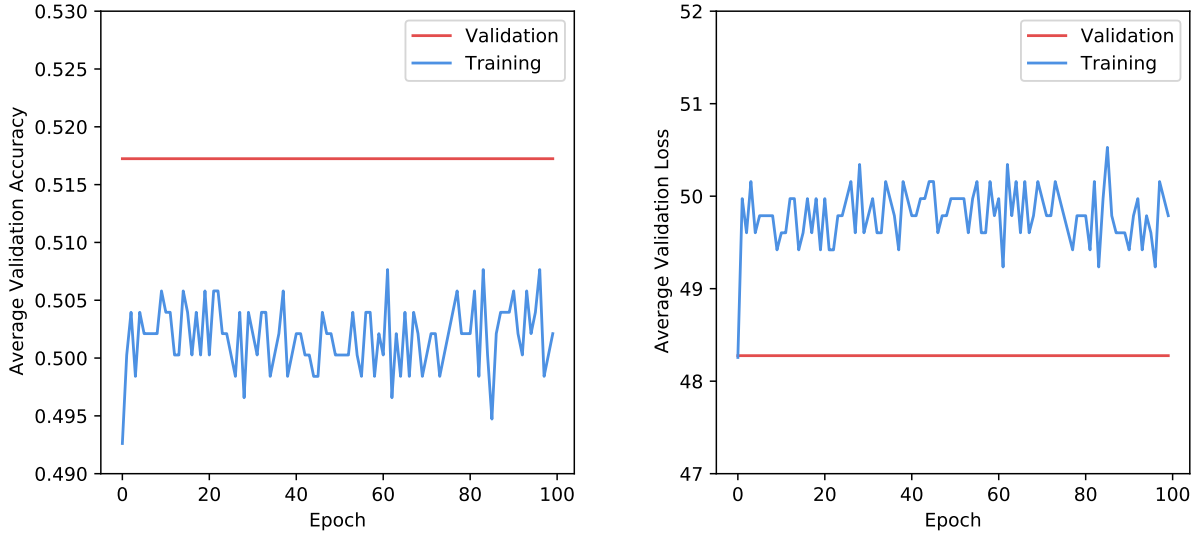


Figure 7: Results of training BagNet with ResNet-50 backbone on the LUAD – LYM dataset with default weight initialization.

4.1.2 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 [5, 39]. In order to verify if this premise stays valid 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 [6]. 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.

These weight initialization approaches were benchmarked by using them for training on the LUAD – LYM binary dataset. Figure 7 shows training and validation results obtained by

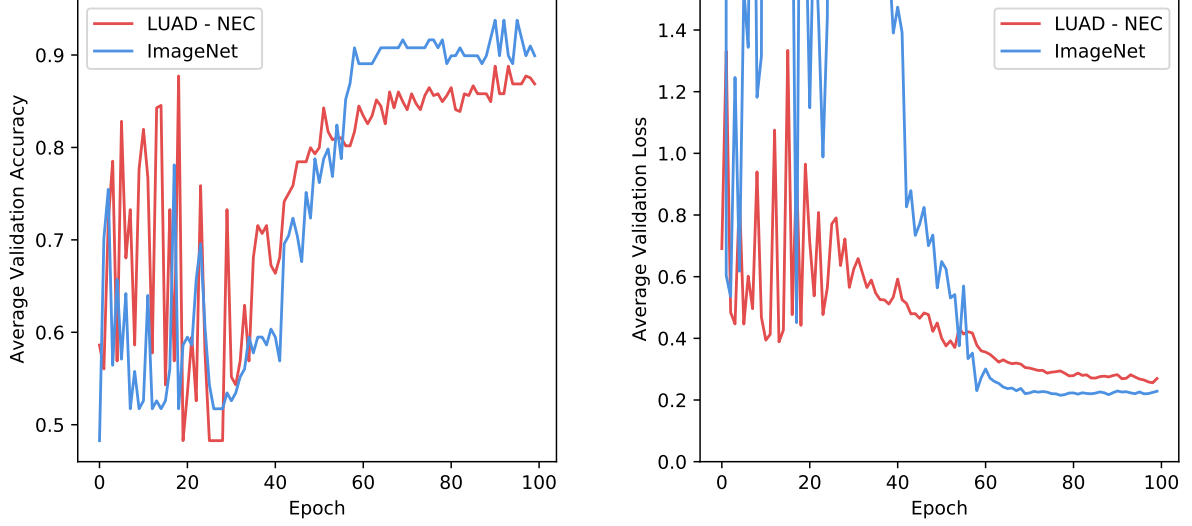


Figure 8: Results of training BagNet with ResNet-50 backbone on the LUAD – LYM dataset with weight initialization using pre-trained models on ImageNet and LUAD – NEC datasets.

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 problem of the model being unable to optimize past a local optimum could possibly be solved by increasing the learning rate.

The results of training the BagNet model with the two different pre-trained sets of weights are shown in Figure 8. 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 [6].

4.1.3 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

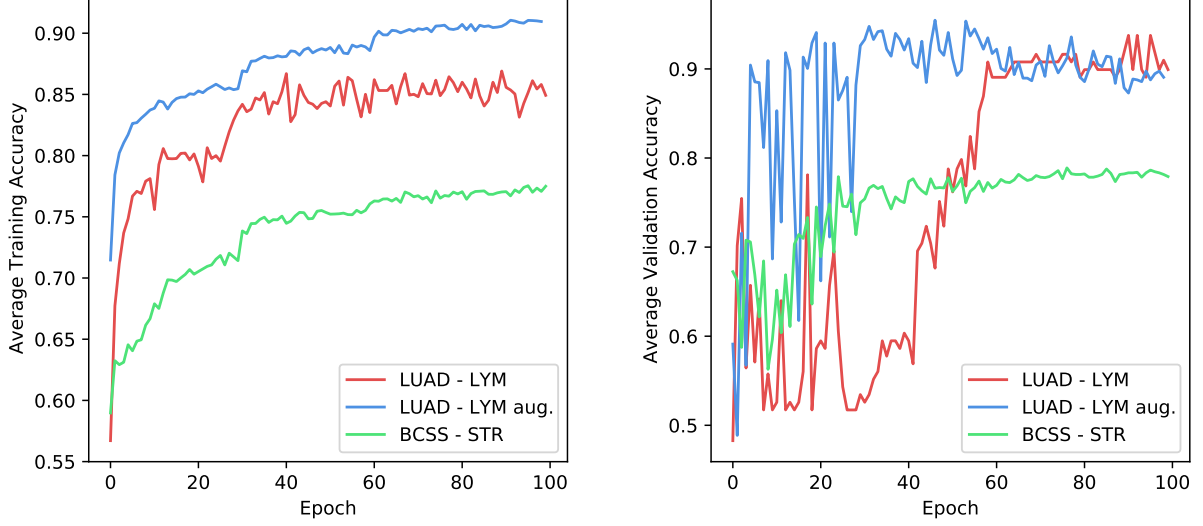


Figure 9: Results of training BagNet with ResNet-50 backbone, ImageNet weight initialization on datasets of different sizes: LUAD – LYM with 1400 training samples and 100 validation samples, augmented LUAD – LYM with 14 000 training samples, and 1000 validation samples, BCSS – STR with 16 000 training samples, and 1600 validation samples.

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 9 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 appears 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 [15].

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.

4.2 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 3, along with the sizes of datasets the models were trained and validated on.

Binary dataset	Avg. val. accuracy	Train set size	Val. set size
LUAD – LYM	0.94	1400	100
LUAD – LYM aug.	0.954	14 000	1000
LUAD – NEC	0.94	6000	40
LUAD – NEC aug.	0.99	60 000	400
LUAD – TE aug.	0.95	80 000	600
BCSS – STR	0.79	16 000	1600

Table 3: Results of training BagNet with ResNet-50 backbone with ImageNet weight initialization on binary datasets of different sizes with and without data augmentation.

By analyzing the accuracy values obtained in this experiment, we can draw 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 a significantly lower accuracy than the models trained on the LUAD datasets. This might be caused by the fact that the BCSS is a less detailed dataset mentioned in Section 4.1.3, a topic that is further discussed in Section 4.3.

By comparing the results obtained on augmented datasets with their non-augmented counterparts in the first four rows of Table 3, 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. Moreover, the results presented in this table demonstrate comparable performances to radiology experts, in most cases even exceeding the capabilities of human pathologists [18], which shows real-world applicability. However, the robustness of the model on WSI-specific artifacts, color variations, or slide deteriorations mentioned in Section 1.3 has not been tested, which might limit model usability in real-world environments.

4.3 Qualitative evaluation

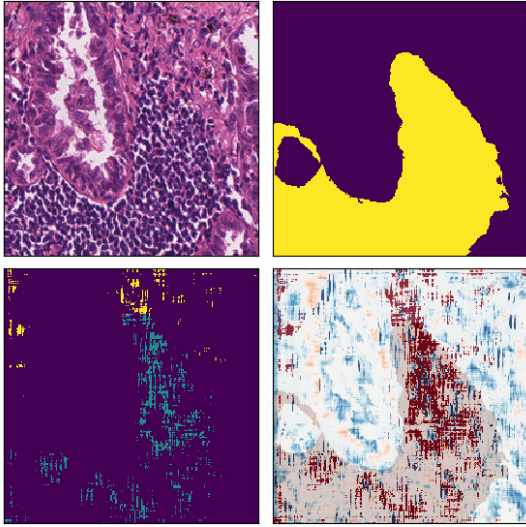
As our approach is intended to replace human specialists in the safety-critical healthcare industry, one of the main reasons for choosing the BagNet architecture was its inherent explainability. To demonstrate expected behavior, we hereby evaluate the decision-making process of our models, which in the case of the adapted BagNet architecture consists of the analysis of the models’ pixel activation heatmaps. The expected behavior is reviewed using the expert annotated segmentation masks included in the validation and test sets of both LUAD and BCSS weakly-annotated datasets, as shown in Figure 5.

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 10 and 11.

Specific examples from datasets of different sizes are shown in Figure 10. 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

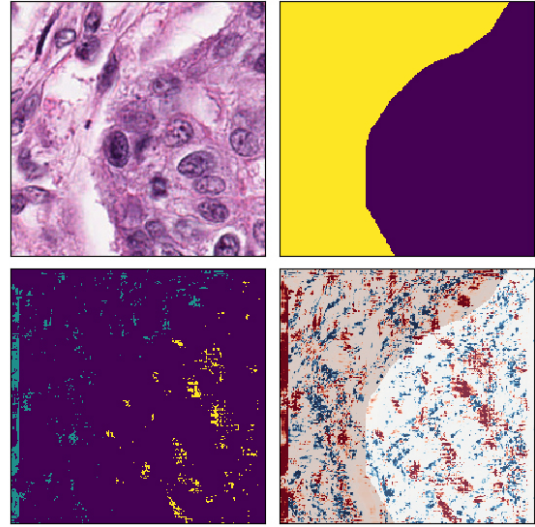
80.38% of the top 5.0% of patches are in the binary mask



LUAD – LYM

without augmentation, 1,400 training samples

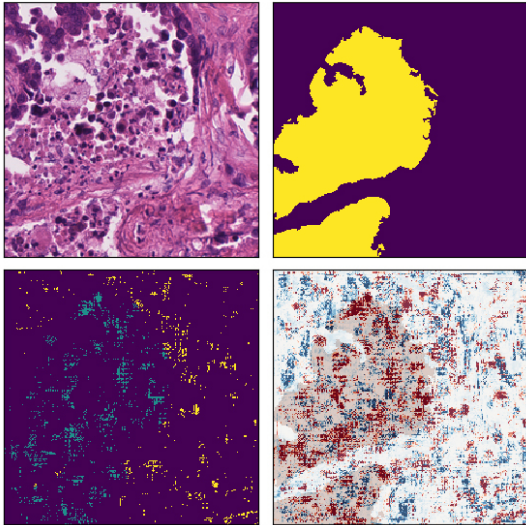
67.26% of the top 5.0% of patches are in the binary mask



BCSS – STR

without augmentation, 16,000 training samples

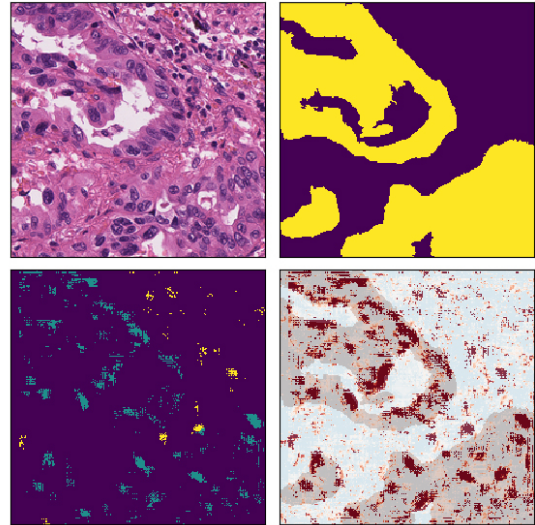
65.55% of the top 5.0% of patches are in the binary mask



LUAD – NEC

with augmentation, 60,000 training samples

89.00% of the top 5.0% of patches are in the binary mask

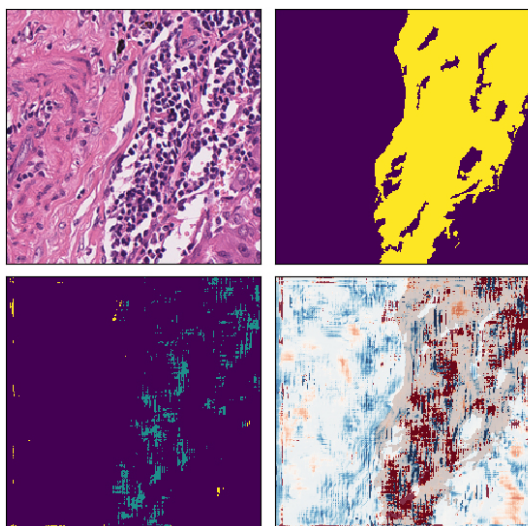


LUAD – TE

with augmentation, 80,000 training samples

Figure 10: Heatmap analysis of BagNet models trained on four different datasets with varying sizes. The performance of each model is illustrated 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 color.

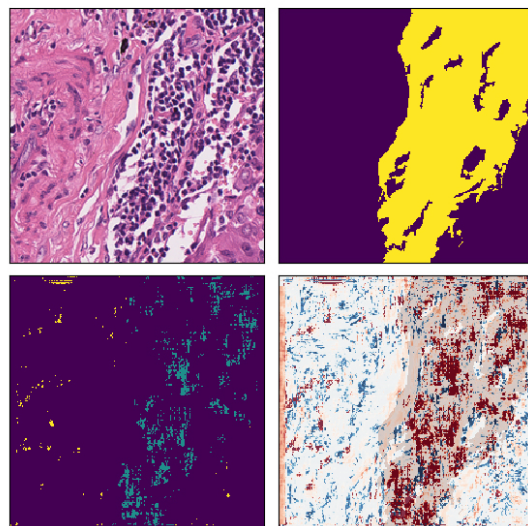
94.10% of the top 5.0% of patches are in the binary mask



LUAD – LYM

without augmentation, 1,400 training samples

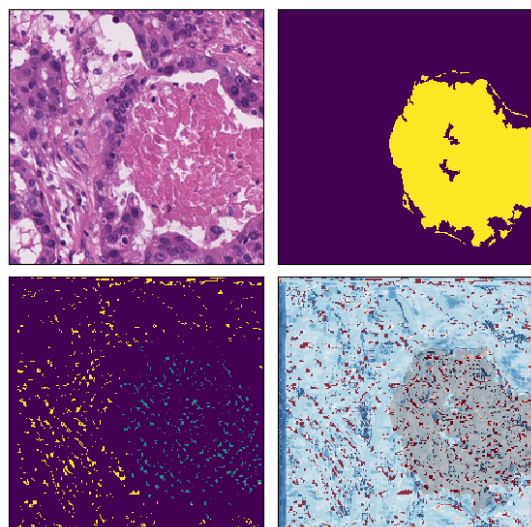
89.59% of the top 5.0% of patches are in the binary mask



LUAD – LYM

with augmentation, 14,000 training samples

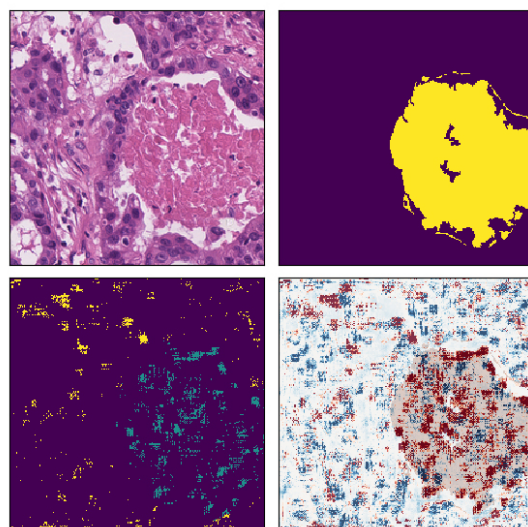
39.00% of the top 5.0% of patches are in the binary mask



LUAD – NEC

without augmentation, 6,000 training samples

62.64% of the top 5.0% of patches are in the binary mask



LUAD – NEC

with augmentation, 60,000 training samples

Figure 11: Heatmap analysis of BagNet models trained for binary classification of two different classes with and without data augmentation. The structure of the illustrations is analogous to the one described in Figure 10.

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.

As mentioned in Section 2.2, model generalizability can be improved by augmenting training data. Therefore, we examine the effects of data augmentation on the decision-making process of the BagNet models by analyzing Figure 11, 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 10, 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 4.2.

5 Conclusion

5.1 Overview

Medical image processing is crucial in modern medicine as imaging technologies become more prevalent. With the continuous digitization of whole-slide image acquisition processes such as tissue sample collection, there is a growing need for specialized algorithms to analyze WSIs accurately and efficiently. Tissue type detection in tumor microenvironments is an example of an important problem in the context of WSI processing, which has a significant impact on cancer initiation and progression. Solving such tasks algorithmically, and thus aiding the work done by human radiologists is becoming feasible due to advancements in fields such as image processing and deep learning.

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.

5.2 Theses

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.

5.3 Discussion

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 [15].

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 [4]. Moreover, model performance could be evaluated on larger validation and testing datasets as well to further solidify results regarding classification accuracy.

Acknowledgements

The authors acknowledge that this work is supported by Babeş–Bolyai University’s Special Scholarship for Scientific Activity, as well as the Collegium Talentum program of the Sapientia Hungariae Foundation. The Márton Áron Vocational College of Eötvös Loránd University also supported the research.

We thank Dr. Bognár Gergő and Dr. Kovács Péter from the Faculty of Numerical Analysis, Eötvös Loránd University for providing the computer hardware used in the experiments.

References

- [1] Khalid AbdulJabbar et al. “Geospatial immune variability illuminates differential evolution of lung adenocarcinoma”. In: *Nature medicine* 26.7 (2020), pages 1054–1062.
- [2] Mohamed Amgad et al. “Structured crowdsourcing enables convolutional segmentation of histology images”. In: *Bioinformatics* 35.18 (2019), pages 3461–3467.
- [3] Christopher J Anders et al. “Understanding patch-based learning of video data by explaining predictions”. In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (2019), pages 297–309.
- [4] Leila Arras, Ahmed Osman, and Wojciech Samek. “CLEVR-XAI: a benchmark dataset for the ground truth evaluation of neural network explanations”. In: *Information Fusion* 81 (2022), pages 14–40.
- [5] Murat Seçkin Ayhan et al. “Clinical validation of saliency maps for understanding deep neural networks in ophthalmology”. In: *Medical Image Analysis* 77 (2022), page 102364.
- [6] Wieland Brendel and Matthias Bethge. “Approximating cnns with bag-of-local-features models works surprisingly well on imagenet”. In: *arXiv preprint arXiv:1904.00760* (2019).
- [7] Saul Calderon-Ramirez et al. “Dealing with scarce labelled data: Semi-supervised deep learning with mix match for covid-19 detection using chest x-ray images”. In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. 2021, pages 5294–5301.
- [8] Renato Cuocolo et al. “Machine learning applications in prostate cancer magnetic resonance imaging”. In: *European radiology experimental* 3.1 (2019), pages 1–8.
- [9] Izak B Dimenstein. “Grossing biopsies: an introduction to general principles and techniques”. In: *Annals of diagnostic pathology* 13.2 (2009), pages 106–113.
- [10] Kunio Doi. “Current status and future potential of computer-aided diagnosis in medical imaging”. In: *The British journal of radiology* 78.suppl_1 (2005), s3–s19.
- [11] Laila El Jiani, Sanaa El Filali, et al. “Overcome medical image data scarcity by data augmentation techniques: A review”. In: *2022 International Conference on Microelectronics (ICM)*. IEEE. 2022, pages 21–24.
- [12] Renato de Filippis et al. “Machine learning techniques in a structural and functional MRI diagnostic approach in schizophrenia: a systematic review”. In: *Neuropsychiatric disease and treatment* (2019), pages 1605–1627.

- [13] Alan G Fraser et al. “The future of cardiovascular imaging and non-invasive diagnosis”. In: *European Journal of Echocardiography* 7.4 (2006), pages 268–273.
- [14] Jiuxiang Gu et al. “Recent advances in convolutional neural networks”. In: *Pattern recognition* 77 (2018), pages 354–377.
- [15] Chu Han et al. “Multi-layer pseudo-supervision for histopathology tissue semantic segmentation using patch-level classification labels”. In: *Medical Image Analysis* 80 (2022).
- [16] Douglas Hanahan and Robert A Weinberg. “Hallmarks of cancer: the next generation”. In: *cell* 144.5 (2011), pages 646–674.
- [17] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pages 770–778.
- [18] Le Hou et al. “Patch-based convolutional neural network for whole slide tissue image classification”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pages 2424–2433.
- [19] Qinghua Huang, Fan Zhang, and Xuelong Li. “Machine learning in ultrasound computer-aided diagnostic systems: a survey”. In: *BioMed research international* 2018 (2018).
- [20] Indu Ilanchezian et al. “Interpretable gender classification from retinal fundus images using BagNets”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III* 24. Springer. 2021, pages 477–487.
- [21] Krzysztof Iniewski. *Medical imaging: principles, detectors, and electronics*. John Wiley & Sons, 2009.
- [22] Neel Kanwal et al. “The Devil is in the Details: Whole Slide Image Acquisition and Processing for Artifacts Detection, Color Variation, and Data Augmentation: A Review”. In: *IEEE Access* 10 (2022), pages 58821–58844.
- [23] Hany Kasban, MAM El-Bendary, and DH Salama. “A comparative study of medical imaging techniques”. In: *International Journal of Information Science and Intelligent System* 4.2 (2015), pages 37–58.
- [24] Vincent S Khoo et al. “Magnetic resonance imaging (MRI): considerations and applications in radiotherapy treatment planning”. In: *Radiotherapy and Oncology* 42.1 (1997), pages 1–15.
- [25] Yong-Ku Kim and Kyoung-Sae Na. “Application of machine learning classification for structural brain MRI in mood disorders: Critical review from a clinical perspective”. In: *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 80 (2018), pages 71–80.

- [26] David K Klassen et al. “Pancreas allograft biopsy: safety of percutaneous biopsy—results of a large experience”. In: *Transplantation* 73.4 (2002), pages 553–555.
- [27] Márton Kolossváry et al. “Advanced atherosclerosis imaging by CT: radiomics, machine learning and deep learning”. In: *Journal of cardiovascular computed tomography* 13.5 (2019), pages 274–280.
- [28] Philippe Lambin et al. “Radiomics: the bridge between medical imaging and personalized medicine”. In: *Nature reviews Clinical oncology* 14.12 (2017), pages 749–762.
- [29] Yann LeCun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pages 2278–2324.
- [30] Xintong Li et al. “A comprehensive review of computer-aided whole-slide image analysis: from datasets to feature extraction, segmentation, classification and detection approaches”. In: *Artificial Intelligence Review* 55.6 (2022), pages 4809–4878.
- [31] Min Lin, Qiang Chen, and Shuicheng Yan. “Network in network”. In: *arXiv preprint arXiv:1312.4400* (2013).
- [32] Xueyan Mei et al. “Artificial intelligence-enabled rapid diagnosis of patients with COVID-19”. In: *Nature medicine* 26.8 (2020), pages 1224–1228.
- [33] Hossein Mohammad-Rahimi et al. “Application of machine learning in diagnosis of COVID-19 through X-ray and CT images: a scoping review”. In: *Frontiers in cardiovascular medicine* 8 (2021), page 638011.
- [34] Sandra Morales, Kjersti Engan, and Valery Naranjo. “Artificial intelligence in computational pathology—challenges and future directions”. In: *Digital Signal Processing* 119 (2021), page 103196.
- [35] Abdullah-Al Nahid, Tariq M Khan, and Yinan Kong. “Hardware implementation of bone fracture detector using fuzzy method along with local normalization technique”. In: *Annals of Data Science* 4 (2017), pages 533–546.
- [36] Sajid Nazir, Diane M Dickson, and Muhammad Usman Akram. “Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks”. In: *Computers in Biology and Medicine* (2023), page 106668.
- [37] Stephen O’Hara and Bruce A Draper. “Introduction to the bag of features paradigm for image classification and retrieval”. In: *arXiv preprint arXiv:1101.3354* (2011).

- [38] Alex M Pagnozzi et al. “A systematic review of structural MRI biomarkers in autism spectrum disorder: A machine learning perspective”. In: *International Journal of Developmental Neuroscience* 71 (2018), pages 68–82.
- [39] Changhyun Park and Heung-Il Suk. “Deep Joint Learning of Pathological Region Localization and Alzheimer’s Disease Diagnosis”. In: *arXiv preprint arXiv:2108.04555* (2021).
- [40] Magdalini Paschali et al. “Deep learning under the microscope: improving the interpretability of medical imaging neural networks”. In: *arXiv preprint arXiv:1904.03127* (2019).
- [41] Tekla S Perry. “Andrew ng x-rays the ai hype”. In: *IEEE Spectrum*, May 2 (2021).
- [42] Daniele Ravi et al. “Degenerative adversarial neuroimage nets: generating images that mimic disease progression”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III* 22. Springer. 2019, pages 164–172.
- [43] Beatriu Reig et al. “Machine learning in breast MRI”. In: *Journal of magnetic resonance imaging* 52.4 (2020), pages 998–1018.
- [44] Eleonora Ruocco et al. “Noninvasive imaging of skin tumors”. In: *Dermatologic surgery* 30 (2004), pages 301–310.
- [45] Olga Russakovsky et al. “Imagenet large scale visual recognition challenge”. In: *International journal of computer vision* 115 (2015), pages 211–252.
- [46] Raul D Santos and Khurram Nasir. “Insights into atherosclerosis from invasive and non-invasive imaging studies: Should we treat subclinical atherosclerosis?” In: *Atherosclerosis* 205.2 (2009), pages 349–356.
- [47] Connor Shorten and Taghi M Khoshgoftaar. “A survey on image data augmentation for deep learning”. In: *Journal of big data* 6.1 (2019), pages 1–48.
- [48] K Kirk Shung, Michael Smith, and Benjamin MW Tsui. *Principles of medical imaging*. Academic Press, 2012.
- [49] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).
- [50] Stanley R Sternberg. “Biomedical image processing”. In: *Computer* 16.01 (1983), pages 22–34.
- [51] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. “One pixel attack for fooling deep neural networks”. In: *IEEE Transactions on Evolutionary Computation* 23.5 (2019), pages 828–841.

- [52] Kim S Suvarna, Christopher Layton, and John D Bancroft. *Bancroft's theory and practice of histological techniques E-Book*. Elsevier health sciences, 2018.
- [53] Christian Szegedy et al. "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pages 1–9.
- [54] Aiham Taleb et al. "3d self-supervised methods for medical imaging". In: *Advances in neural information processing systems* 33 (2020), pages 18158–18172.
- [55] Christian Tesche and Hunter N Gray. "Machine learning and deep neural networks applications in coronary flow assessment: the case of computed tomography fractional flow reserve". In: *Journal of Thoracic Imaging* 35 (2020), S66–S71.
- [56] Erico Tjoa and Cuntai Guan. "A survey on explainable artificial intelligence (xai): Toward medical xai". In: *IEEE transactions on neural networks and learning systems* 32.11 (2020), pages 4793–4813.
- [57] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. "Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge". In: *Contemporary Oncology/Współczesna Onkologia* 2015.1 (2015), pages 68–77.
- [58] Bas HM Van der Velden et al. "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis". In: *Medical Image Analysis* (2022), page 102470.
- [59] Chunhao Wang et al. "Artificial intelligence in radiotherapy treatment planning: present and future". In: *Technology in cancer research & treatment* 18 (2019), page 1533033819873922.
- [60] Mingliang Wang et al. "Multi-task exclusive relationship learning for Alzheimer's disease progression prediction with longitudinal data". In: *Medical image analysis* 53 (2019), pages 111–122.
- [61] Anthony B Wolbarst. *Looking within: how X-ray, CT, MRI, ultrasound, and other medical images are created, and how they help physicians save lives*. Univ of California Press, 1999.
- [62] Guang-Quan Zhou and Yong-Ping Zheng. "Automatic fascicle length estimation on muscle ultrasound images with an orientation-sensitive segmentation". In: *IEEE Transactions on Biomedical Engineering* 62.12 (2015), pages 2828–2836.
- [63] Guang-Quan Zhou et al. "Automatic measurement of spine curvature on 3-D ultrasound volume projection image with phase features". In: *IEEE transactions on medical imaging* 36.6 (2017), pages 1250–1262.