From CNNs to Vision Transformers – A Comprehensive Evaluation of Deep Learning Models for Histopathology

Maximilian Springenberg, Annika Frommholz, Markus Wenzel, Eva Weicken, Jackie Ma, and Nils Strodthoff

Abstract—While machine learning is currently transforming the field of histopathology, the domain lacks a comprehensive evaluation of state-of-the-art models based on essential but complementary quality requirements beyond a mere classification accuracy. In order to fill this gap, we conducted an extensive evaluation by benchmarking a wide range of classification models, including recent vision transformers, convolutional neural networks and hybrid models comprising transformer and convolutional models. We thoroughly tested the models on five widely used histopathology datasets containing whole slide images of breast, gastric, and colorectal cancer and developed a novel approach using an image-to-image translation model to assess the robustness of a cancer classification model against stain variations. Further, we extended existing interpretability methods to previously unstudied models and systematically reveal insights of the models’ classification strategies that allow for plausibility checks and systematic comparisons. The study resulted in specific model recommendations for practitioners as well as putting forward a general methodology to quantify a model’s quality according to complementary requirements that can be transferred to future model architectures.

Index Terms—breast, gastrointestinal tract, computer-aided detection and diagnosis, end-to-end learning in medical imaging, evaluation and performance, machine learning, microscopy, molecular and cellular imaging, neural network, pattern recognition and classification, segmentation

I. INTRODUCTION

MACHINE learning (ML) has the potential to transform the field of histopathology, where expert pathologists visually examine stained tissue specimen under the microscope, e.g., for cancer diagnosis [1]. New ML technologies make the rapid, automatic analysis of large numbers of digitized whole slide images (WSIs) possible and promise to alleviate the burden of the time-consuming examination by human experts in traditional workflows. Automatic image analysis bears also the potential to enhance these workflows with quantitative metrics, e.g., by accurately quantifying tumor infiltrating lymphocytes across entire WSIs instead of relying on a coarse visual estimate [2]. These innovative improvements are mainly driven by the steady progress of deep learning (DL), which is an ML sub-discipline focusing on multi-layer neural networks. Latest ML methods show very promising results on a broad range of analysis tasks in histopathology (see [3]–[7] for recent reviews) and already reach performance levels of human pathologists for specific tasks [8].

The predictive performance of an ML model on a given task is typically assessed with standard metrics such as accuracy or sensitivity/specificity. However, additional quality criteria, such as the robustness and interpretability of the model, are highly relevant for the clinical application as well, and are complementary to the quantitative evaluation of the predictive performance. The term robustness has many facets ranging from robustness against shifts in the data distribution, to robustness against input perturbations or adversarial attacks. Concerning interpretability, the field of explainable artificial intelligence (XAI) has witnessed tremendous advances in the past few years and lead to the development of methods that allow certain insights into the decision process of complex ML models (see [9]–[14] for reviews). XAI techniques have already been adopted in the domain of histopathology to a certain degree (see [15] for a recent review and [16] for a best-practice paper).

In this paper, we aim to quantitatively compare different state-of-the-art model architectures in the context of histopathology from a comprehensive point of view with respect to predictive performance, interpretability and robustness. On five histopathology datasets, we benchmark convolutional neural networks (CNNs) but also vision transformers (ViTs) and hybrid convolutional-transformer models, where especially ViTs and hybrid models have not been explored extensively yet. Concerning interpretability, we provide relevance (or: attribution) heatmaps for all models with the layer-wise relevance propagation (LRP) framework, and are thereby extending this framework to previously unstudied model architectures. In addition, we propose a quantitative evaluation scheme to assess the plausibility of the resulting relevance heatmaps. Finally, we focus on the robustness of the models with respect to stain variations. For a quantitative assessment, we propose to use an image-to-image translation model (CycleGAN) in order to disentangle the effects of stain variation from other sources of distribution shifts.

Our main contributions are (a) performance assessment of state-of-the-art convolutional models, vision transformers and hybrids across five publicly accessible datasets covering different cancer and tissue types, (b) implementation of LRP for all considered models and quantitative evaluation of the resulting heatmaps across the whole dataset through combination with segmentation models, and (c) robustness evaluation.
with respect to cross-stain variation through a learned image-to-image translation model.

II. MATERIALS AND METHODS

A. Datasets and Tasks

Five openly accessible datasets serve for the evaluation of the models regarding predictive performance, interpretability, and robustness. The datasets are briefly characterized in Table I. Most notably, the datasets cover different cancer entities, such as breast cancer (a) metastasis in lymph-nodes (PCam [17]–[19]) (b) tissue (BreaKHis [1], IDC [20]–[22]), gastric cancer (GasHisSDB [23]), and colorectal cancer (MHIST [24]). Targeted towards identifying invasive ductal carcinoma (IDC) breast cancer tissue, the IDC dataset used in this work replicates the patched IDC dataset from [20] at a higher spatial resolution using information from the original data sources [21], [22]. With a sliding window approach, we join adjacent patches from [20] creating non-overlapping 100 × 100 patches out of the original 50 × 50 patches filling missing patches by the corresponding parts from the original source [21], [22]. We mostly focus on breast cancer throughout this manuscript and use the GasHisSDB and the MHIST datasets to demonstrate that our findings are not specific for breast cancer and can be adapted to other cancer entities. Most of the datasets that we included here employ a patch-wise binary classification task, namely “benign vs. malignant” (PCam, IDC, GasHis, MHIST). Only BreaKHis distinguishes different benign and malignant subtypes and is therefore framed as a multi-class classification task. All datasets had been labeled by expert pathologists.

| Dataset               | Cancer type                        | Classes | Training | Valid. | Test | Total |
|-----------------------|------------------------------------|---------|----------|--------|------|-------|
| PCam                  | Breast (metastasis)                | 2       | 262,144  | 32,768 | 32,768| 327,680|
| BreaKHis              | Breast                             | 8       | 1,005    | 504    | 504  | 2,013 |
| IDC                   | Breast                             | 2       | 26,734   | 10,009 | 16,410 | 32,768 |
| GasHisSDB             | Gastric                            | 2       | 13,313   | 6,657  | 13,134 | 33,284 |
| MHIST                 | Colorectal                         | 2       | 2,175    | -      | 977  | 3,152 |

B. Models

A wide selection of state-of-the-art models with different working principles is included in this study. ResNet50 [25] and Inception V3 [26] serve as representatives for modern convolutional neural networks (CNNs). Vision transformers (ViTs) are beginning to challenge CNNs in computer vision for natural images, but have so far received only limited attention in the field of histopathology; see [27] for first results in this direction. We consider both the original ViT [28] as well as a ViT with a convolutional stem [29] which we henceforth refer to as ViTc, as different architecture categories, throughout this work. It is worth stressing that even though this latter model architecture is often referred to as a hybrid CNN-ViT-model, it is more a ResNet model with non-local attention layers rather than an actual transformer model such as the ViT or the ViTc.

Finally, we also consider hybrid models, including the GasHis-Transformer, composed of a BoTNet50 and an Inception V3 model, which showed promising results in [27], as well as selected further combinations of two single models each. The notion of an hybrid model in the context of this work is that of two model architectures, whose penultimate layers, containing the feature vectors, are concatenated and fed into a fully connected classification layer, benefiting from diverse feature extraction. In the following, we will refer to hybrids of two models \( m_1 \) and \( m_2 \) as \( m_1+m_2 \). We feature the architectures of ResNet50, BoTNet50, Inception V3, ViT, and ViT with convolutional stem (ViTc) as single models, with Inception V3 being the strongest performing model across validation splits. With the intent to avoid overinflation, we limited this work to the hybrids of Inception V3 and ResNet50 (IV3+RN50), BoTNet50 (IV3+BN50, i.e. GasHis), ViT (IV3+ViT), as well as ViTc (IV3+ViTc), motivated by the performance of Inception V3 across validation splits. To pursue a fair comparison between architectures, we will distinguish between single and hybrid models, treating them as different architecture categories, throughout this work.

Training was conducted with reproducibility in mind. All models were trained from scratch for the same number of epochs, using Adam as optimizer and the same data augmentations: Random hue shifts spanning the entire hue spectrum, horizontal, as well as vertical flips and rotations by up to 180° were included in the augmentation procedure. The model with the maximal area under the curve (AUC) of the receiver operating characteristic on the validation split at the end of an epoch was selected, for datasets that were furnished with a validation split (PCam, BreaKHis, IDC, GasHisSDB; see Table I). We release the code for training and evaluation of all models in a corresponding code repository [31]. The models differ in their complexity, which is influenced by the number of parameters (ResNet50: 24.6M, BoTNet50: 18.8M, Inception V3: 24.3M, ViT: 11.4 M, ViTc: 96.1M), but also by a range of several further factors [22], which might play a decisive role when choosing a model in practice.

C. Quality aspects and their evaluation

1) Predictive performance: The performance of the models (see Section II-B) in solving the binary-/multi-classification problems associated with the five datasets (see Section II-A Table I) was assessed with the common performance metrics of accuracy and AUC. As a particular methodological strength of this study, we address two major sources of uncertainty in our assessment, and consider both the uncertainty due to the randomness of training process and the uncertainty due to the finiteness of the test set, see Section II-A for details.

2) Interpretability: To gain insights into the opaque decision making of the complex DL models, we draw on the growing body of research literature in the field of XAI, see [11], [14], [33] for reviews. More specifically, we use LRP to calculate relevance heatmaps which highlight regions in input
space that speak for or against the classification decision. The LRP-implementations follow [34] for ResNet and [35], [36] for transformers. We implemented LRP rules for the BoTNet, the Inception model as well as the hybrid models along the same lines, none of which have been discussed in the literature so far. This provides a coherent framework for the comparison of heatmaps between different models. To investigate the models’ strategies for cancer classification, we propagated the label using LRP for all cancer positive samples of the PCam dataset. As we aimed to identify image regions that are deemed relevant for cancer detection by the model, we averaged over color channels and set all negative relevance to zero, referring to this approach as $R_{mean,max}$ pooling. This procedure results in a relevance map emphasizing image regions that are relevant for cancer-positive classification according to LRP.

As one of the main contributions in this work, we put forward a way to quantify relevance maps in terms of the fraction of relevance attributed to certain semantically meaningful subsections of the original image. Here, we distinguish nuclei, background and tissue surrounding nuclei. Nuclei segmentation maps were obtained from an nnU-Net [37] that was trained on nuclei segmentation data MoNuSeg [38]. The background was segmented with gray value thresholding and subsequent morphological opening. Pixels that were neither nuclei nor background were considered as tissue surrounding the nuclei. We then studied the overlap of relevant input areas with the nuclei, tissue, or background segmentation maps. This approach enables us to quantitatively compare different models with respect to their decision making strategies. Potentially, we can find evidence for models using different (or comparable) strategies to achieve a different (or comparable) accuracy. To some degree, the approach also makes it possible to assess the plausibility of the decision making.

Cancer is characterized by rapid cell division activity, which requires first and foremost DNA replication processes in the cell nuclei, and then nuclear division, visible under the microscope. Tumor proliferation rate/speed is therefore an important biomarker for tumor grading and as a prognostic factor. It is commonly assessed by pathologists counting the number of mitotic nuclei in hematoxylin & eosin (H&E) stained histological slides under the microscope and is expressed by the “mitotic activity index” (MAI, number of mitoses in 2 mm² tissue area) [39], [40]. Mitosis describes the process of cell nucleus division across several phases, from the division of the genetic material to the strangulation of the cell body [41]. For automatic cancer detection, it would therefore be a useful strategy if the ML models were considering cell nuclei in particular.

We accordingly hypothesized that a plausible decision of a model, which classifies tissue as cancerous or not, should accordingly be attributed to a considerable degree to image areas where cell nuclei are located. Otherwise, if the decision of the model can be attributed mostly to other image areas outside of the nuclei or even to the background, we consider this decision as less plausible, and recommend further checks for potential confounding factors that may influence the decision making. Overlap between relevance heatmaps and the segmentation maps (of nuclei, tissue, and background) was quantified with the point biserial correlation coefficient [42]–[44] (equivalent to Pearson correlation), and with the metric of mass accuracy that is the relevance in the segmentation mask divided by the entire relevance [45].

3) Robustness: Staining of the tissue specimen varies from lab to lab depending on the precise experimental protocol (amount and effective duration of chemical exposure, waiting periods, etc.) and the used scanner equipment [46], which can result in subtle hue or contrast shifts in the WSIs. Cancer classification models that are planned to be employed in practice should be able to generalize across labs and should be robust against staining variations. We assessed the robustness as follows. First, we trained classification models either on the original BreaKHis or IDC train split. (To allow a direct comparison between both datasets, we trained the binary classifier on BreaKHis using the IDC subclass as positive and all benign samples as negative class.) Then, we evaluated the resulting classifiers separately on the original BreaKHis/IDC test splits and on BreaKHis/IDC images that were stain transformed with a separate image-to-image translation model. This approach (see Figure 1) enabled us to to disentangle the effect of stain variation from the actual image content of the tissue slices.

Cycle-consistent adversarial networks (CycleGANs) represent a way to train image-to-image translation models based on unpaired data [47]. An adaptation of the CycleGAN generator structure by [48] has shown to improve the preservation of structural features in stain transformation applications. Similar to fully convolutional encoder-decoder architectures, the image is downsampled to a low-dimensional representation and then upsampled again, trained to retain the relevant features. The generator is an nnU-Net with several skip connections added to learn relevant structural information. Other GAN implementations that can be used for stain normalization are StainGAN [39] and RestainNet [46]. All of these implementations aim to perform a template color matching approach in order to adjust the color distribution of an image or dataset to a specified target while retaining image contents.

In our CycleGAN, the generator also has a generic nnU-Net structure [37], which is an adaptation of the U-Net...
specifically for medical image segmentation. The generator output is calculated through a hard tanh function as the final activation function resulting in output values between 0 and 1. We hypothesize that the internal connections from the nnUNet force the model to retain the high-resolution structural components and primarily learn the color transformation.

The discriminator architecture consists of three convolutions followed by a 2D mean pooling layer which reduces the dimensionality of the discriminator output by 1. Each of the convolutions is followed by a 2D batch normalization and an activation function. Firstly and secondly, a ReLU activation is used and the third convolution utilizes a sigmoid function. The following 2D averaging results in the discriminator relying on color information rather than structural information when comparing original to synthesized images.

We trained a CycleGAN for 150 epochs on BreaKHis and a subset of the IDC dataset. Data were preprocessed by resizing to a size of 96 × 96 and by random vertical and horizontal flips (p = 0.5). The learning rate was scheduled to start decaying after half of the training was complete. The generator and discriminator learning rates were set to 0.0002 and the Adam optimizer (β1 = 0.9, β2 = 0.999) was used for training. Mean squared error served as the loss. The trained CycleGAN provides two generator models, \( G_{\text{BreaKHis} \rightarrow \text{IDC}} \) and \( G_{\text{IDC} \rightarrow \text{BreaKHis}} \), for transformations in both directions that enabled us to transform BreaKHis images to mimic the IDC distribution and vice versa. We created images with a fake IDC staining from the original BreaKHis test split, and images with a fake BreaKHis staining from the original IDC test split.

A realistic, meaningful assessment of cross-distribution robustness can only be achieved if characteristics including tissue type, magnification, and image resolution are similar. We adjusted all factors as far as possible for similarity, and used the BreaKHis dataset with an optical magnification of \( 96 \times 96 \) and the IDC dataset for the robustness validation, because of the similarity in both tissue type and magnification.

### III. RESULTS AND DISCUSSION

#### A. Predictive performance

The predictive performance of the selected set of models on the five datasets is summarized in Table I. Inception V3 is the best-performing single model with only very few exceptions and for both performance metrics (accuracy, AUC). The two related hybrid models GasHis(I3+BN50) and I3+RN50 turned out as overall best-performing architectures.

In most cases, Inception V3 is followed by the other convolution models, where BoTNet50 does not establish a clear advantage through its multi-head self-attention layers compared to its purely convolutional counterpart ResNet50. On most datasets, these models are followed by the vision transformers, where again the vision transformer with convolutional stem shows an advantage compared to the vanilla vision transformer, most likely through its larger inductive bias, which is beneficial on the comparably small histopathology datasets. In terms of datasets, MHIST stands out as the only dataset where the Inception V3 model falls back behind the other convolutional architectures and where the vision transformer with convolutional stem performs best as opposed to a very poorly performing vanilla vision transformer. The example of the even smaller BreaKHist dataset shows that this behavior cannot entirely be attributed to the small dataset size and deserves future investigations. We investigated additional hybrid models formed by combining the best-performing single-model Inception V3 with one of the other single models.

Our results may speak against the hypothesis that the global context extracted via the BoTNet50 is a key ingredient for the performance of GasHis [27], because we find that the combination of Inception V3 with a standard ResNet50 (i.e. a

### TABLE II

| Model                  | Acc. [%] | AUC          |
|------------------------|----------|--------------|
| **I3**                 |          |              |
| ResNet50               | 85.70 ± 0.51 | 0.9355 ± 0.0080 |
| BoTNet50               | 83.35 ± 2.11 | 0.9094 ± 0.0196 |
| Inception V3           | 87.36 ± 0.57 | 0.9431 ± 0.0046 |
| ViT                    | 78.42 ± 0.42 | 0.8800 ± 0.0072 |
| ViTc                   | 86.40 ± 0.29 | 0.9442 ± 0.0016 |
| GasHist (I3+BN50)      | 87.86 ± 0.26 | 0.9464 ± 0.0022 |
| IV3+RN50               | 87.80 ± 0.21 | 0.9469 ± 0.0052 |
| IV3+ViT                | 85.43 ± 0.60 | 0.9436 ± 0.0033 |
| IV3+ViTc               | 87.43 ± 0.40 | 0.9427 ± 0.0058 |
| MHIST (×40)            |          |              |
| ResNet50               | 76.74 ± 0.78 | 0.9477 ± 0.0067 |
| BoTNet50               | 78.81 ± 1.90 | 0.9639 ± 0.0056 |
| Inception V3           | 91.29 ± 0.66 | 0.9911 ± 0.0009 |
| ViT                    | 83.58 ± 0.62 | 0.9734 ± 0.0026 |
| ViTc                   | 75.07 ± 2.34 | 0.9440 ± 0.0091 |
| GasHist (I3+BN50)      | 90.34 ± 0.48 | 0.9511 ± 0.0011 |
| IV3+RN50               | 91.37 ± 0.46 | 0.9911 ± 0.0001 |
| IV3+ViT                | 83.70 ± 0.86 | 0.9726 ± 0.0021 |
| IV3+ViTc               | 85.29 ± 1.44 | 0.9791 ± 0.0013 |

**CODE**

```python
# Import necessary libraries
import torch
import torch.nn as nn
import torch.optim as optim
from torch.utils.data import Dataset, DataLoader

# Define the discriminator architecture
class Discriminator(nn.Module):
    def __init__(self):
        super(Discriminator, self).__init__()
        self.layers = nn.Sequential(
            nn.Conv2d(in_channels=3, out_channels=64, kernel_size=4, stride=2, padding=1),
            nn.LeakyReLU(0.2, inplace=True),
            nn.MaxPool2d(kernel_size=2, stride=2),
            nn.Conv2d(in_channels=64, out_channels=128, kernel_size=4, stride=2, padding=1),
            nn.LeakyReLU(0.2, inplace=True),
            nn.MaxPool2d(kernel_size=2, stride=2),
            nn.Conv2d(in_channels=128, out_channels=256, kernel_size=4, stride=2, padding=1),
            nn.LeakyReLU(0.2, inplace=True),
            nn.MaxPool2d(kernel_size=2, stride=2),
            nn.Conv2d(in_channels=256, out_channels=512, kernel_size=4, stride=2, padding=1),
            nn.LeakyReLU(0.2, inplace=True),
            nn.MaxPool2d(kernel_size=2, stride=2),
            nn.Conv2d(in_channels=512, out_channels=1024, kernel_size=4, stride=2, padding=1),
            nn.LeakyReLU(0.2, inplace=True),
            nn.MaxPool2d(kernel_size=2, stride=2),
            nn.Conv2d(in_channels=1024, out_channels=1, kernel_size=4, stride=1, padding=0),
            nn.ReLU(inplace=True)
        )

    def forward(self, x):
        return self.layers(x)

# Define the generator architecture
class Generator(nn.Module):
    def __init__(self):
        super(Generator, self).__init__()
        self.layers = nn.Sequential(
            nn.Conv2d(in_channels=100, out_channels=512, kernel_size=4, stride=2, padding=1),
            nn.ReLU(inplace=True),
            nn.Conv2d(in_channels=512, out_channels=256, kernel_size=4, stride=2, padding=1),
            nn.ReLU(inplace=True),
            nn.Conv2d(in_channels=256, out_channels=128, kernel_size=4, stride=2, padding=1),
            nn.ReLU(inplace=True),
            nn.Conv2d(in_channels=128, out_channels=64, kernel_size=4, stride=2, padding=1),
            nn.ReLU(inplace=True),
            nn.Conv2d(in_channels=64, out_channels=3, kernel_size=4, stride=2, padding=1),
            nn.Tanh()
        )

    def forward(self, z):
        return self.layers(z)

# Define the dataset class
class FashionMNISTDataset(Dataset):
    def __init__(self, data, transform=None):
        self.data = data
        self.transform = transform

    def __len__(self):
        return len(self.data)

    def __getitem__(self, index):
        image, label = self.data[index]
        if self.transform:
            image = self.transform(image)
        return image, label

# Define the data loaders
train_loader = DataLoader(dataset=train_dataset, batch_size=64, shuffle=True)
valid_loader = DataLoader(dataset=valid_dataset, batch_size=64, shuffle=False)
```
combination of two convolutional models) yields a comparable performance on all investigated datasets. The more likely explanation for the GasHis performance is that the combination of two different models add a bit of diversity to the combined model that consequently leads to (mostly) small performance gains compared to a single Inception V3 model.

To address the uncertainty due to the randomness of the training process, we performed \( k = 5 \) training runs for each architecture and report empirical mean as well as standard error over the resulting set of trained models in Table [II] The statistical uncertainty of the performance results due to the finiteness of the test set was quantified via empirical bootstrapping on the test set (100 iterations). For statistical comparison of two model architectures, we combine both uncertainties following [50]. We claim a set of models \( M_1 \) to perform not significantly worse than another set of models \( M_2 \) if in at least \( \frac{k(k+1)}{2k^2} = 60\% \) of the \( k \times k \) direct comparisons between the runs of both architectures, the 95\% confidence intervals for the respective score differences do overlap with zero or \( M_1 \) performs better than \( M_2 \). The threshold is induced by the convention that a set of trained models \( M = \{m_i | 1 \leq i \leq k \} \) shall not be significantly worse than \( M \) itself, even if the bootstrapping results in an ordering \( m_1 \prec m_2 \prec \ldots \prec m_k \), where \( m_i \prec m_j \) denotes that \( m_i \) performs worse than \( m_j \) (i.e. the 95\% confidence intervals obtained via bootstrapping for the respective score differences do not overlap with zero). Comparisons that resulted in statistically significant differences are indicated with asterisks and diamonds in Table [II] To summarize, there are consistent trends in the predictive performance that are stable across both metrics and all datasets, which still favor convolutional as opposed to transformer architectures. Our analysis suggests using Inception V3, the best-performing single model, as default architecture for the exploration of new histopathology datasets, notwithstanding the small performance gains that can be achieved by combining it with other models into hybrid models. Furthermore, we have gathered evidence that the presented models are, in principle, well suited for classifying a variety of cancer types, ranging from breast- (PCam, BreaKHis, IDC), gastric (GasHisSDB) to colorectal (MHIST) cancer.

B. Interpretablity

1) Relevant image segments: The relevance distributions across nuclei, surrounding tissue, and background, quantitatively characterize the model behaviour (see Table [III] The classification models appear to attribute relevance mostly to nuclei segments, while surrounding tissue and background seem to be negligible. In other words, nuclei are the most important image segments for the classification decision.

The relevance heatmaps of all models gravitate towards cell nuclei segments with a mass accuracy above the random threshold in every case. Pearson correlation between heatmaps and nuclei segments was slightly but significantly \( (p < 0.05) \) positive for GasHis, Inception V3 and ViT. On the one hand, this finding aligns very well with the observed strong predictive performance, in the case of GasHis and Inception V3 (two best-performing models on PCam). On the other hand, there are also top-performing models such as the IV3+RN50 hybrid model and the decently performing ViT model that show a comparably low mass accuracy in the nuclei category. Obviously, performance does not solely depend on the focus of the model on the most appropriate image regions but also on the way the information from these regions is combined inside the model, an insight which is not conveyed through relevance heatmaps.

| Model              | Mass acc. | Pearson r | Pearson p |
|--------------------|-----------|-----------|-----------|
| **NUCLEI**         |           |           |           |
| ResNet50           | 0.50      | 0.052     | 0.056     |
| BoTNet50           | 0.48      | 0.028     | 0.064     |
| Inception V3       | 0.53      | 0.102     | 0.010     |
| ViT                | 0.51      | 0.065     | 0.029     |
| ViTc               | 0.47      | 0.012     | 0.067     |
| GasHis(IV3+BN50)   | **0.56**  | 0.12      | 0.006     |
| IV3+RN50           | 0.49      | 0.031     | 0.023     |
| IV3+ViT            | 0.54      | 0.169     | 0.023     |
| IV3+ViTc           | 0.51      | 0.093     | 0.028     |
| Random             | 0.47      | **0.0**   | 0.05      |
| **TISSUE**         |           |           |           |
| ResNet50           | 0.40      | -0.037    | 0.059     |
| BoTNet50           | 0.42      | 0.008     | 0.074     |
| Inception V3       | 0.38      | -0.039    | 0.013     |
| ViT                | 0.39      | -0.044    | 0.033     |
| ViTc               | 0.42      | -0.001    | 0.073     |
| GasHis(IV3+BN50)   | 0.33      | -0.110    | 0.007     |
| IV3+RN50           | 0.40      | -0.029    | 0.023     |
| IV3+ViT            | 0.36      | -0.134    | 0.027     |
| IV3+ViTc           | 0.38      | -0.085    | 0.029     |
| Random             | 0.42      | 0         | >0.05     |
| **BACKGROUND**     |           |           |           |
| ResNet50           | 0.10      | -0.006    | 0.097     |
| BoTNet50           | 0.10      | -0.027    | 0.101     |
| Inception V3       | 0.08      | -0.039    | 0.032     |
| ViT                | 0.10      | -0.001    | 0.069     |
| ViTc               | 0.11      | -0.003    | 0.108     |
| GasHis(IV3+BN50)   | **0.12**  | -0.007    | 0.039     |
| IV3+RN50           | 0.11      | 0.029     | 0.030     |
| IV3+ViT            | 0.10      | -0.018    | 0.095     |
| IV3+ViTc           | 0.11      | 0.012     | 0.067     |
| Random             | 0.11      | 0         | >0.05     |

In comparison to the Inception V3 model, the other single models ResNet50, BotNet50 and ViT, which performed comparably poor, show a sizable gap in terms of mass accuracy. Thus, models that are stronger in terms of predictive performance tend to show a higher correlation of relevance heatmaps with nuclei segmentation masks (as measured with mass accuracy and Pearson correlation).

It is an interesting observation that even though the predictive performance of Inception V3 is comparable to GasHis (no statistically significant difference in terms of AUC) on PCam, there are pronounced differences in terms of the distribution of relevance onto the different image segments (nuclei, surrounding tissue, background), most notably in the relative amount of relevance that is attributed to the surrounding tissue (mass accuracy of 0.38 for Inception V3 vs. 0.33 for
GasHis; see Table III. This suggests that there are different strategies that can lead to comparable predictive accuracy. As a further remark, the focus on nuclei reflects the accordance with strategies used by human pathologists (cf. Section II-C2). More in-depth interpretability might shed light on alternative classification strategies that are potentially followed by certain ML models. In any case, our analysis provides the first step towards a quantitative evaluation of interpretability methods in the context of histopathology.

It can be observed that the focus on nuclei is not uniformly distributed across the segmented nuclei, e.g., in the examples in Figure 2. Relevance is clustered around individual subsegments, i.e., nuclei, rather than the segmented area of joint nuclei as a whole. Thus, the models distribute high relevance on specific nuclei, while not all nuclei are relevant to the models. This observation aligns with the fact that we cannot distinguish between healthy and cancerous segments of nuclei, considering that the segmentation model was trained on generic nuclei.

In general, all models show very little positive to negative correlation with the segmentation masks of tissue and background. GasHis and Inception V3 show a significant \((p < 0.05)\) negative Pearson correlation of relevance heatmaps with tissue surrounding the nuclei and, respectively, with the background. This result confirms the above finding and our hypothesis that strong predictive models indeed focus on cell nuclei and less on surrounding tissue or background, for their classification decision.

2) ViT attention per head: Attention-based architectures such as the ViT have the advantage that the architecture-inherent attention maps provide additional insights into the model behavior. Here, we study the attention maps obtained from the CLS-token of the topmost layer. Individual ViT heads seem to distribute attention to specific segments and appear to focus mostly on the nuclei and background segments, rather than on tissue (see Table IV).

Due to the ability of the ViT’s final layer to join attention maps, the model may even divide its attention into multiple heads. Furthermore, a head does not have to focus on one thematic segment, that is to say it may deal with certain aspects of different segment classes at once, which can be observed in Figure 3 and Table IV. Our methodology provides a quantitative way to assess earlier qualitative findings on the specialization of attention heads in the literature [51].

To summarize, interpretability methods make it possible to gain insights into the strategies pursued by different models to reach the classification decision. Our main contribution in this respect is correlating the corresponding heatmaps with the output of segmentation models to assess quantitatively and beyond single examples how much relevance the models put on different image compartments. Whereas all models put most relevance on nuclei, there are hints for distinct model strategies that can lead to a comparable predictive performance.

C. Robustness

1) Re-stained test data via CycleGAN: In an effort to assess the robustness of the models against stain variations, we first generated re-stained test images with a CycleGAN (see Section II-C3) and present the re-staining results in this section before turning to the results of the actual robustness tests using these re-stained images in the next Section III-C2.

Exemplary modified versus original images from BreaKHis/IDC are displayed in Figure 3. These images stem from the test splits of both datasets and were chosen to represent a variety of original images from the distributions. To not only rely on a qualitative review of the generated samples, we compared histograms from the CycleGAN-created test sets to the target distribution. Every 10 epochs,
a model was used to modify the two test sets and a hue-value-histogram was calculated from the results. We chose to display the hue-value of the HSV color space as function values in the histograms to condense the color information into one value for an easier comparison (instead of three values in the case of the RGB color space). From the histogram evolution, we chose the 150th epoch as a suitable stopping point for the CycleGAN training. The hue value histograms from both transformation directions are shown in Figure 5. The original starting distribution and the target distribution are added to the diagram, for reference, along with the final distribution after CycleGAN training.

Using the CycleGAN’s generator models, \( G_{IDC \rightarrow \text{BreakHis}} \) and \( G_{\text{BreakHis} \rightarrow IDC} \), we were able to map the patches staining from one datasets distribution to the other and vice versa, such that they mimic another. However small deviations remain, as can be observed from Figure 5. The color distribution at large hue values is unimodal for IDC whereas Fake IDC shows two peaks as in the source dataset. Therefore the CycleGAN is not to be considered as a stain normalizer, for it does not perfectly match the distribution of colors, but rather as a model to mimic a cross-distribution staining. We consider the generators to produce re-stained patches that are more realistic in aesthetic than global transformation methods, such as shifts in hue and saturation, due to their ability to adjust colors locally with respect to visual features. The generators ability to attain local color transformation can also be observed from the examples displayed in Figure 4.

2) Robustness evaluation: In order to single out the specific effect of staining variations between labs, we tested the predictive performance of the classification models on the novel cross-stained, but in-distribution images (see previous Section III-C1), and compared the test results with the common setup of train/test-data originating from the same (in-stain/in-distribution) or different sites (cross-stained/cross-

Fig. 4. Example images created by the CycleGAN color transformation. Columns 1 and 3 show original images from the IDC and BreakHis test split. The columns aside are created by the CycleGAN from the original images and imitate the staining of the respective other dataset.

Fig. 5. Hue value histograms for the BreakHis and IDC datasets and the stain transformations in both directions created by a CycleGAN. The blue dotted curve shows the starting hue value distribution for the respective original dataset, the orange dashed curve the resulting hue value distribution at the end of the CycleGAN training for 150 epochs, and the black solid curve the respective target distribution.
distribution. Table V lists the results of our robustness evaluation; compare also the illustration in Figure 1. The models consistently perform best on in-stain/in-distribution, worse on cross-stain/in-distribution and worst on cross-stain/cross-distribution samples, which matches our expectations. (Only ViTc forms an exception with similar results in both cross-stain cases when trained on BreaKHis.) The performance of models trained on IDC and tested on ‘Fake BreaKHis’ images (that had been re-stained with the generator GIIC→BreaKHis) drops for about 3.6% to 7.7% in comparison to the in-stain/in-distribution test (IDC), and for 25.2% to 40.2% in the corresponding cross-stained/cross-distribution test (BreaKHis). When models trained on BreaKHis were tested on images with a ‘Fake IDC’ staining (generated with G_BreaKHis→IDC), the performance drops roughly for 13.4% to 20.4% due to the cross-staining, and for about 14.6% to 48.9% in the corresponding cross-distribution case (IDC).

Thus, the specific effect of staining variations between different labs, which we simulated using a CycleGAN, already does pose a substantial challenge for the cancer classification models examined (see the performance drops in percent in the column “cross-stain, in-dist.” in Table V). Nevertheless, the effect of differing distributions between train and test data appears to be the dominating factor and results in a further large performance decrease (column “cross-stain, cross-dist.” in Table V). This effect occurs even though we tried to mitigate it as much as possible through weighted sampling of the normal and abnormal classes to match the proportions of the respective original in-distribution test sets.

In light of the results, we come to the sobering observation, that it is hard to identify architectures that are particularly robust in either scenario. No architecture is truly robust to challenges posed by cross-stain and cross-distribution data. The results reflect the brittleness of deep learning methods in general. Despite the successes in suppressing the proneness of differentiable models to overfitting in recent years, the training dataset still governs the models ability to generalize on out-of-distribution data. Robustness is not a property that can be taken for granted, but requires dedicated efforts to be achieved and incorporated into the state of the art.

We would like to highlight that, for a sincere evaluation of the classifiers’ robustness, we had included global hue value shifts in our data augmentation during training (see section II-B), inducing a robustness to global color transformations. As a result, classification performance indeed did not significantly change compared to the results from Table II when evaluating on test data to which different global color transformations had been applied (data not shown). The proposed robustness test utilizing the CycleGAN-generated, re-stained patches provides more realistic local color transformations and makes it possible to disentangle the cross-distribution- from the stain-variation-effect.

The effects of cross-stain- and cross-distribution-data could be lessened by adapting to distribution shifts in the training data, such as including the respective generators GIIC→BreaKHis, GIIC→IDC into augmentation or through other domain adaptation techniques [32]–[34], though this is beyond the scope of this work, since we focused on evaluating trained classifiers. In particular, our results suggest that the commonly used global color transformations do not lead to enhanced robustness against more realistic stain transformations.

In summary, we propose to use generative models to assess the robustness against stain variations. Based on our results we cannot single out specific architectures in terms of a particularly high robustness but see a general necessity for further research in this direction.

### IV. Conclusion

Relating back to the main achievements brought forward in the introduction, we summarize the three main findings of our study as follows. (a) In a thorough benchmarking study involving a broad range of state-of-the-art image recognition models across five public histopathology datasets, we identify a common convolutional architecture – Inception V3 – as best-performing model, which still outperforms transformer-based models. (b) Our quantitative assessment of attribution (heat)maps reveals that all models focus on nuclei. Slight differences in the relevance distribution can be seen as an indication for different classification strategies, which can be exploited, e.g., through the formation of ensemble models. (c) Our proposed methodology allows to quantitatively assess the robustness against staining differences and revealed, irrespective of the model architecture, insufficient robustness (see Section III-C2) as a major obstacle for the application of deep-learning-based image recognition models in the wild.

### ACKNOWLEDGMENT

The authors would like to thank Angel Cruz-Roa and Anant Madabhushi for helpful correspondence on the IDC dataset.
[37] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation,” Nature methods, vol. 18, no. 2, pp. 203–211, 2021. https://doi.org/10.1038/s41592-020-01008-z

[38] N. Kumar, R. Verma, S. Sharma, S. Bhargava, A. Vahadane, and A. Sethi, “A dataset and a technique for generalized nuclear segmentation for computational pathology,” IEEE Transactions on Medical Imaging, vol. 36, no. 7, pp. 1550–1560, 2017.

[39] A. Sohail, A. Khan, N. Wahab, A. Zameer, and S. Khan, “A multi-phase deep CNN based mitosis detection framework for breast cancer histopathological images,” Scientific Reports, vol. 11, no. 1, pp. 1–18, 2021. https://doi.org/10.1038/s41598-021-85652-1

[40] M. Veta, P. J. Van Diest, M. JiwA, S. Al-Janabi, and J. P. Pluim, “Mitosis counting in breast cancer: Object-level interobserver agreement and comparison to an automatic method,” PloS one, vol. 11, no. 8, p. e0161286, 2016. https://doi.org/10.1371/journal.pone.0161286

[41] J. R. McIntosh and M. F. Koonce, “Mitosis,” Science, vol. 246, no. 4930, pp. 622–628, 1989. https://doi.org/10.1126/science.2683078

[42] L. Arras, A. Osman, and W. Samek, “Ground truth evaluation of neural network explanations with clevr-xai,” arXiv preprint arXiv:2003.07258, 2020. https://arxiv.org/abs/2003.07258

[43] B. Zhao, J. Lin, C. Liang, Z. Yi, X. Chen, B. Li et al., “RestainNet: a self-supervised digital re-stainer for stain normalization,” 2022. https://arxiv.org/pdf/2202.13804

[44] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2242–2251. https://doi.org/10.1109/ICCV.2017.244

[45] L. Arras, A. Osman, and W. Samek, “Ground truth evaluation of neural network explanations with clevr-xai,” arXiv preprint arXiv:2003.07258, 2020. https://arxiv.org/abs/2003.07258

[46] T. de Bel, J.-M. Bokhorst, J. van der Laak, and G. Litjens, “Residual cyclegan for robust domain transformation of histopathological tissue slides,” Medical Image Analysis, vol. 70, p. 102004, 2021. https://doi.org/10.1016/j.media.2021.102004

[47] M. T. Shaban, C. Baur, N. Navab, and S. Albarqouni, “StainGAN: Stain style transfer for digital histological images,” 2018. https://arxiv.org/abs/1804.01601

[48] T. Mehari and N. Strodthoff, “Self-supervised representation learning without target labels,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 3, pp. 766–785, 2021. https://doi.org/10.1109/TPAMI.2019.2945942

[49] R. J. Chen and R. G. Krishnan, “Self-supervised vision transformers learn visual concepts in histopathology,” in NeurIPS 2021 workshop on Learning Meaningful Representations of Life, 2021.

[50] M. Wang and W. Deng, “Deep visual domain adaptation: A survey,” Neurocomputing, vol. 312, pp. 135–153, 2018. https://www.sciencedirect.com/science/article/pii/S0925231218306684

[51] H. Guan and M. Liu, “Domain adaptation for medical image analysis: A survey,” IEEE Transactions on Biomedical Engineering, vol. 69, no. 3, pp. 1173–1185, 2022. https://doi.org/10.1109/TBME.2021.3117407

[52] W. M. Kouw and M. Loog, “A review of domain adaptation without target labels,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 3, pp. 766–785, 2021. https://doi.org/10.1109/TPAMI.2019.2945942