Computer Methods and Programs in Biomedicine 232 (2023) 107441

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Improved classification of colorectal polyps on histopathological images with ensemble learning and stain normalization

Sena Busra Yengec-Tasdemir[a,b,*], Zafer Aydin[b,c], Ebru Akay[d], Serkan Dogan[e], Bulent Yilmaz[f,b]

[a] School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast, BT39DT, United Kingdom
[b] Department of Electrical and Computer Engineering, Abdullah Gul University, Kayseri, 38080, Turkey
[c] Department of Computer Engineering, Abdullah Gul University, Kayseri, 38080, Turkey
[d] Pathology Clinic, Kayseri City Hospital, Kayseri, 38080, Turkey
[e] Gastroenterology Clinic, Kayseri City Hospital, Kayseri, 38080, Turkey
[f] Department of Electrical Engineering, Gulf University for Science and Technology, Mishref, 40005, Kuwait

ARTICLE INFO

Article history:
Received 14 January 2023
Revised 5 February 2023
Accepted 21 February 2023

MSC:
41A05
41A10
65D05
65D17
Colorectal Polyps
Colonic Polyp Classification
Histopathology Image Classification
Computer-aided Diagnosis
Clinical Decision Support System
Ensemble of Deep Convolutional Neural Networks
ConvNeXt
Transfer Learning

ABSTRACT

Background and Objective: Early detection of colon adenomatous polyps is critically important because correct detection of it significantly reduces the potential of developing colon cancers in the future. The key challenge in the detection of adenomatous polyps is differentiating it from its visually similar counterpart, non-adenomatous tissues. Currently, it solely depends on the experience of the pathologist. To assist the pathologists, the objective of this work is to provide a novel non-knowledge-based Clinical Decision Support System (CDSS) for improved detection of adenomatous polyps on colon histopathology images.

Methods: The domain shift problem arises when the train and test data are coming from different distributions of diverse settings and unequal color levels. This problem, which can be tackled by stain normalization techniques, restricts the machine learning models to attain higher classification accuracies. In this work, the proposed method integrates stain normalization techniques with ensemble of competitively accurate, scalable and robust variants of CNNs, ConvNexts. The improvement is empirically analyzed for five widely employed stain normalization techniques. The classification performance of the proposed method is evaluated on three datasets comprising more than 10k colon histopathology images.

Results: The comprehensive experiments demonstrate that the proposed method outperforms the state-of-the-art deep convolutional neural network based models by attaining 95% classification accuracy on the curated dataset, and 91.1% and 90% on EBHI and UniToPatho public datasets, respectively.

Conclusions: These results show that the proposed method can accurately classify colon adenomatous polyps on histopathology images. It retains remarkable performance scores even for different datasets coming from different distributions. This indicates that the model has a notable generalization ability.

1. Introduction

According to global cancer statistics published in 2021 [1] colorectal cancer (CRC) is one of the most common causes of cancer deaths. Most CRC cases develop as an adenomatous polyp. If adenomatous polyps are detected early, the CRC can be prevented by the removal of the colonic adenomatous polyps. They can be seen during a colonoscopy procedure. During the colonoscopy, polyps are found and removed for histopathological examination, which provides the diagnostic information for treatment and long-term follow-up. Therefore, the detection of adenomatous polyps on histopathological images is critically important for patients. However, currently, this solely depends on the expert pathologists' experience. This study proposes a decision support system for them in the hope to reduce the missed detection of malicious polyps.

In the clinical workflow of polyp classification, a key diagnostic challenge is the differentiation of adenomatous polyps from non-adenomatous tissues. The adenomatous polyp types are tubular, villous, and tubulovillous adenomas. Moreover, adenomatous polyps have the potential to develop into cancer,

* Corresponding author.
E-mail addresses: sena.yengec@agu.edu.tr, s.yengectasdemir@qub.ac.uk (S.B. Yengec-Tasdemir), bulent.yilmaz@agu.edu.tr (B. Yilmaz).

https://doi.org/10.1016/j.cmpb.2023.107441
0169-2607/

while hyperplastic (i.e. non-adenomatous or normal) polyps are usually not likely to show malignancy potential. Therefore, distinguishing adenomatous/neoplastic polyp tissue from non-adenomatous/hyperplastic/normal tissue is a significant step in cancer screening.

Since early diagnosis is vital, there is a growing demand for cancer screening programs. As the demand for screening increases, the workload of pathologists increases, and consequently it gets harder and harder to detect disease at an early stage. In order to carry out this process faster and more accurately, Clinical Decision Support System (CDSS) can be employed, which can ease this labour-intensive work and minimize the mistakes of the traditional approaches. In this work, a CDSS is proposed to assist experts by providing the classes of each histopathology image and highlighting the most suspected areas with Grad-Cam method.

The steady advancement of Deep Learning (DL) based approaches brought a great interest in medical image classification tasks. There are extensive studies conducted with numerous methods focusing on the individual diagnosis of colorectal cancer on histopathology images, such as classification of colorectal adenocarcinoma [2–14], colon polyp classification [4,5,15–23].

In [16], Korbar et al. proposed a methodology to classify colon polyp types on the colonic whole-slide images (WSIs). In this work, WSIs are cropped into patches, then the extracted patches are classified with a ResNet architecture. This methodology achieved an accuracy of 93%. In [15] Korbar et al. proposed a new framework, following their previous work, to visualize attention map on the WSIs. Wei et al. proposed a hierarchical framework to extend the classification of the tissue patches to the whole slide. Moreover, Wei et al. developed a curriculum learning scheme to classify polyps [21]. The proposed methodology achieved an area-under-curve (AUC) of 88.2%. Song et al. [17] emphasized the importance of using different patch sizes for grading and classification of colonic adenomas. Nasir-Moin et al. [19] developed an artificial intelligence (AI) augmented digital system for polyp classification. Performance of the proposed methodology was evaluated by using 238 external slides. Perlo et al. [20] introduced a methodology to grade the dysplasia of the colorectal polyps. Moreover, in their work they compared the model performance on gray scale, RGB and Macenko stain normalized histology images. In [8] Sarwinda et al. employed a ResNet architecture to classify colorectal WSIs into malignant and benign. In their work, they employed Contrast Limited Adaptive Histogram Equalization (CLAHE) methodology as an image pre-processing technique. Recently, Bilal et al. [2] proposed a framework to classify colonic WSIs as neoplastic or normal with a weakly supervised deep learning. They evaluated the performance of their system by using custom collected WSIs and The Cancer Genome Atlas (TCGA) database, and an overall area under the receiver operating characteristic curve (AUROC) of 0.9746 was achieved. Yildirim et al. [12] suggested a CNN based network for detection of the colon cancer on WSI. Zhou et al. [13] employed global labels to localize the cancerous regions on the colonic WSIs. Three experts evaluated the performance of the suggested methodology. Similarly, Gupta et al. [4] used a customized Inception-ResNet-v2 Type 5 (IR-v2 Type 5) model for classification and localization of the abnormal tissues from colonic WSIs. In another work, Ho et al. [5] classified colonic biopsy WSIs as high risk and low risk with an AUC of 91.7%. Moreover, Wang et. al [22] proposed a model that used unsupervised contrastive learning, and achieved an accuracy score of 64.29% on UniToPatho database. Following their work, Wang et al. suggested clustering-guided contrastive learning for polyp classification task and achieved an accuracy score of 66.55% on UniToPatho database [23]. However, most of the state-of-the-art deep learning models for colonic polyp classification employs a single deep CNN algorithm.

Ensemble learning methods are widely used in medical image classification tasks, such as polyp classification on colonoscopy images [24,25], colon cancer detection on histopathology images [14,26,27], breast cancer detection on histopathology images [28]. In [29], Kumar et al. proposed a model that ensembled AlexNet and LeNet architectures, which achieved a greater accuracy than the AlexNet and LeNet architectures alone. Kallipolitis et al. [7] ensembled EfficientNet variants to detect cancer from pathology images of breast and colon.

Ensemble methods are widely employed in breast cancer classification tasks on histopathology images. They can make more robust decisions since they employ various classifiers as base learners, which can capture different levels of information contained in the latent features. However, they have not been used in colonic adenomatous polyp detection on histology images. To the best of our knowledge, this study is the first to use ensemble methods to classify colonic histological images as adenomatous and non-adenomatous. The main contributions of this study are as follows:

- In this study, we explore state-of-the-art pre-trained Deep CNN algorithms' performances on our custom dataset. To the best of our knowledge, this study is the first to comprehensively evaluate widely used stain normalization techniques namely, Stain-GAN, Stain-Net, Vahandane, Macenko and Reinhard by combining with state-of-the art Deep CNN models for classification of adenomatous and non-adenomatous colonic polyp tissues.
- This study is one of the first studies which employs ConvNeXt architecture on colon histopathology images for polyp classification task.
- We propose a model which ensembles the pre-trained ConvNeXt-Tiny and ConvNeXt-Base variants to classify adenomatous and non-adenomatous tissues on colonic histopathology images. Moreover, the variants are further tailored to the problem by network modifications at the image representation levels. In order to comprehensively evaluate and assess the generalizability of the proposed model, we also employ publicly available UniToPatho and EBHI databases [30,31]. The proposed ensemble model achieves an accuracy of 95% on our custom dataset.
- Additionally, in order to ensure the explainability of the proposed model, the Grad-Cam method is used. The attention map of the model is explored for adenomatous and non-adenomatous images. We believe that these Grad-Cam visual outputs can help pathologists to see and judge the decision process of the model.

The paper is organised as follows: The second section *Materials and Methods* includes explanations about the data collection process, stain normalization, ensemble methods, ConvNeXt arcitecture and the proposed model. The following section (Section 3) explains the experimental setup and the settings of the tests that are performed. Section 4 demonstrates the comprehensive test results. Finally, the paper is concluded with a discussion section (Section 5). The complementary details about the test results are added to the Appendix section (Section Appendix A).

2. Material and methods

2.1. Data collection

The histological slides used in this study were collected from 84 patients who underwent colorectal cancer screening since May 2018 at Kayseri City Hospital, Kayseri, Turkey[1] Forty-six of the 84

[1] This study was approved by Kayseri City Hospital Ethics Committee and Erciyes University Clinical Research Ethics Committee.

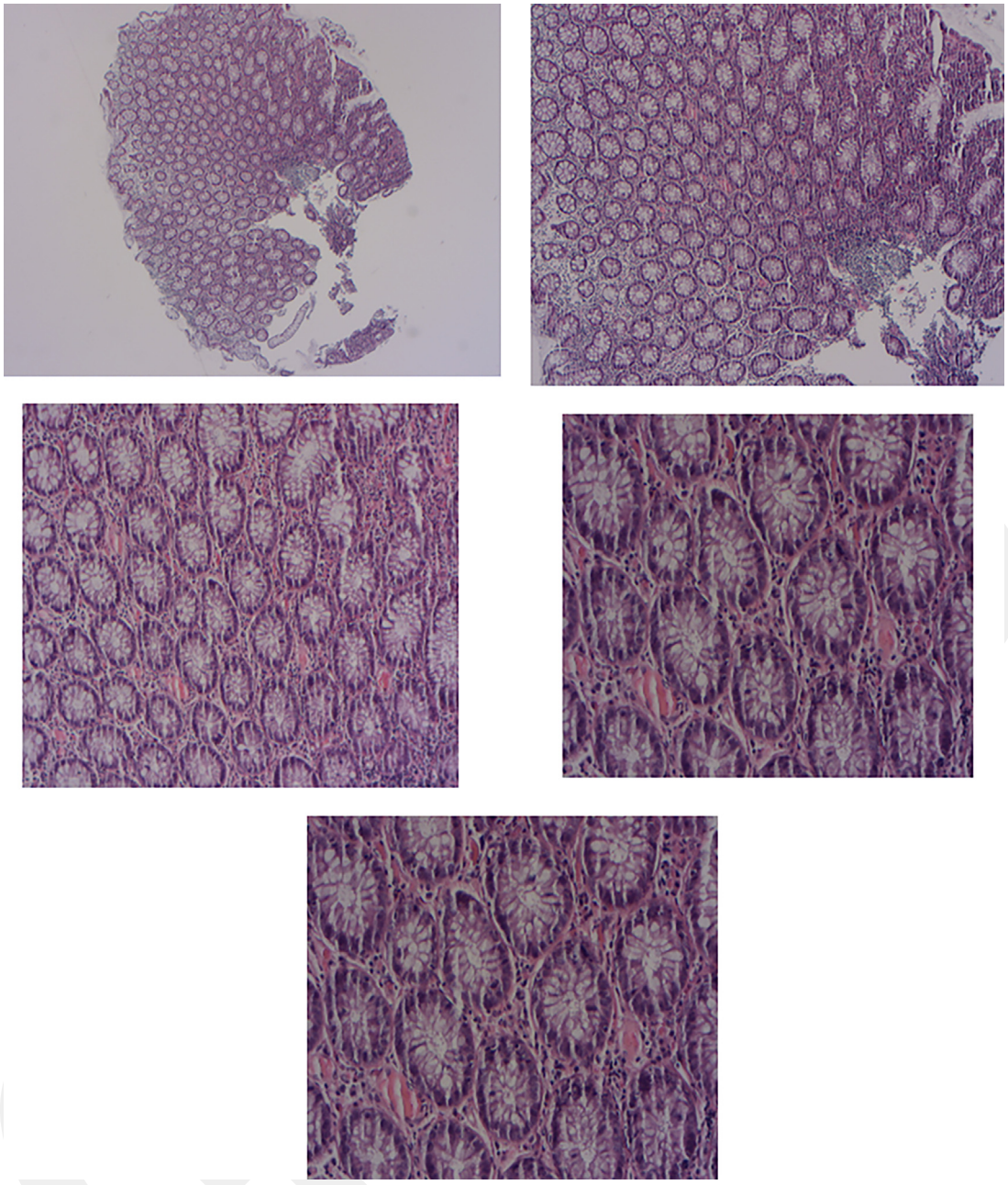

Fig. 1. Histology images of an adenomatous sample tissue for different magnification levels.

patients are male, while the other 38 are female. The age range is between 19 and 89 and the average age is 62.

The dataset contains samples of both adenomatous polyp and non-adenomatous tissues for each patient. The tissue samples were examined under microscopy with 5 different magnification levels and each magnification level is included in the custom dataset. Samples of different magnification settings can be seen in Fig. 1. The magnification levels are: x2.5 (Fig. 1a), x5 (Fig. 1b), x10 (Fig. 1c), x20 (Fig. 1d), and x40 (Fig. 1e). In total, 671 slides were collected, 359 of which belong to adenomatous polyps and 312 slides belong to non-adenomatous tissues (hyperplastic polyps, normal tissue and chronic inflammation). The detailed labelling of the collected samples was done by two expert pathologists for each slide and magnification level.

Four hundred seventy (470) slides were randomly selected for the training, 101 for the validation and 100 for the test set. The train, test and validation sets were separated patient based. That is, there were no whole-slide images that belong to the same patient in two different sets. A detailed description of the collected slides can be found in Table 1.

Table 1
Number of samples and patients for each class in our dataset.

Class	Number of Samples	Number of Patients
Adenoma	359	52
Hyperplasia	181	29
Normal/Chronic Inflammation	130	30

In order to assess the generalizability of the models and stain normalization techniques, randomly selected 152 slides were ac-

quired from publicly available UniToPatho and EBHI databases. UniToPatho database contains 9536 hematoxylin and eosin stained patches extracted from 292 whole-slide images, where each of the slides has a magnification of 20 ×. The WSIs belong to the following classes: normal tissue, hyperplastic polyp, tubular adenoma and tubulo-villous adenoma [30]. EBHI is composed of 5532 WSIs which have the following categories, normal, low-grade and high-grade intra-epithelial neoplasm, and adenocarcinoma, divided into four magnifications of 40×, 100×, 200× and 400× [31].

In most cases, an adenomatous slide may contain one or more different adenomatous tissue structures alongside a normal tissue structure. Usually, an expert pathologist decides whether a WSI contains adenomatous tissue or not by looking through the whole image and the structures in it. Therefore, in this study, each of the WSI is individually classified.

2.2. Stain normalization

Deep CNN algorithms have a great capacity to fit a dataset with high precision. However, this precision challenges the model to generalize for the unseen data. Moreover, if there is a domain shift in training and testing data, the model must be robust and reliable for real-world scenarios.

The domain shift problem is commonly encountered between different WSIs. This difference can be the consequence of different staining protocols, slide preparation or medical center, etc [32]. For instance, in Fig. 2 it can be seen that the color intensities of two different WSIs are quite different. This affects the quality of the trained model. To this end, to ensure good generalization ability, a deep CNN algorithm should be robust to domain shifts. In order to address this issue, different stain normalization techniques are proposed by the researchers. In the literature, the following stain normalization techniques are widely used: Vahandane, Macenko, Reinhard, Stain-GAN, Stain-Net. Vahandane, Macenko and Reinhard methods are more traditional techniques, while Stain-GAN and Stain-Net models are based on Generative Adversarial Network (GAN) structures [33–37]. Fig. 3 shows the outputs of different stain normalization techniques applied on an adenomatous image sample from our custom dataset.

2.3. Ensemble learning

Ensemble learning is a technique which combines various classifiers to enhance classification performance. Different classifiers can capture different information and therefore, ensemble classifiers may result in better accuracy as compared to base learners. Furthermore, ensemble learning methods are widely used in different medical image classification tasks [38]. In [29], Kumar et al. suggested that different CNN classifiers can learn various levels of semantic image representation. In that work, AlexNet and LeNet architectures are fine-tuned on medical images. The proposed method achieved greater accuracy than the AlexNet and LeNet architectures alone.

2.4. Convnext architecture

ConvNeXt architecture has recently been proposed by Liu et al. [39]. This architecture takes the advantage of both the attention-based classifiers and traditional ResNet architectures to compete with the performance of Vision Transformers (ViTs). ConvNeXt architecture is motivated to capture global dependencies by large receptive fields and utilizes convolutions with large kernels as the main building block [40]. Moreover, ConvNeXt is a pure CNN architecture, that can outperform the Swin Transformer for ImageNet-1K classification [39]. The architecture of a ConvNeXt block is presented in Fig. 4 and the ConvNeXt architecture is shown in Fig. 5.

Table 2
Different configurations of the ConvNeXt variants.

Model / Configurations	Number of Channels (C) in each stage	Number of Blocks (B) in each stage
ConvNeXt-T	(96, 192, 384, 768)	(3, 3, 9, 3)
ConvNeXt-S	(96, 192, 384, 768),	(3, 3, 27, 3)
ConvNeXt-B	(128, 256, 512, 1024)	(3, 3, 27, 3)
ConvNeXt-L	(192, 384, 768, 1536)	(3, 3, 27, 3)
ConvNeXt-XL	(256, 512, 1024, 2048)	(3, 3, 27, 3)

To propose the ConvNeXt architecture, in their work, Liu et al. modernized a ResNet architecture by gradually incorporating the essence of the Swin Transformers into the network. Typically, the network starts with a stem cell in which input images are processed. The stem cell implementation of the ConvNeXt architecture is composed of a "Patchify" design which implements a patchify layer using a 4×4, stride 4 convolutional layers.

The following stages of the network are composed of ConvNeXt blocks. In each stage, the number of blocks has a ratio of 3:3:9:3. A ConvNeXt block contains a depth-wise convolution which is followed by 1×1 convolutions. The depth-wise convolution implements a special type of group-wise convolution by grouping the channels. The combination of depth-wise convolution and 1×1 convolutions performs a similar effect to a property that is shared between vision transformers. Additionally, ConvNeXt architecture implements a Gaussian Error Linear Unit (GELU) as an activation function between the two 1×1 convolution layers and uses layer normalization instead of batch normalization. Furthermore, various ConvNeXt variants are suggested, namely, ConvNext-Tiny (T), -Small (S), -Base (B), -Large (L) and -X-Large (XL). The diversity of the variants differs as the number of channels and the number of blocks changes for each stage. Table 2 shows the different configurations of the variants.

2.5. The proposed ensemble of convnexts framework

The proposed framework to classify colonic histological images as either adenomatous or non-adenomatous is shown in Fig. 6. Previous studies on colonic polyp classification problems predominantly employ a single deep CNN algorithm. According to previous studies, [41–,44] CNN architectures play an important role for the classifier performance. As stated in the [45], deep residual CNN architectures are used for more complex problems, while shallower CNNs are used for simple problems. Additionally, ensemble methods perform better than a single deep CNN algorithm, because its base classifiers can interpret various properties of an input image. Consequently, we designed the proposed framework by employing an ensemble of ConvNeXt variants. We selected the ConvNeXt architecture since it is more suitable to classify adenomatous colonic WSIs than the attention-based networks or regular CNNs because of the following reasons:

Unlike ViT [46], it does not require a large amount of data in training. Since there are limited number of samples in our custom dataset, this makes ConvNeXt more convenient than data-hungry attention-based methods such as ViT. Moreover, in contrast to CNNs, it can capture longer dependencies because of its large receptive field. Similar to what experts do, it can recognize an adenomatous polyp structure in a histopathological image by visually inspecting spatially distant cell structures. This makes ConvNeXt more suitable for this task, while it is challenging for CNNs to capture those distant correlations.

The base classifiers of the ensemble model are ConvNeXt-Tiny and ConvNeXt-Base models, which are pre-trained on ImageNet-21k dataset. In order to make those networks more suitable to

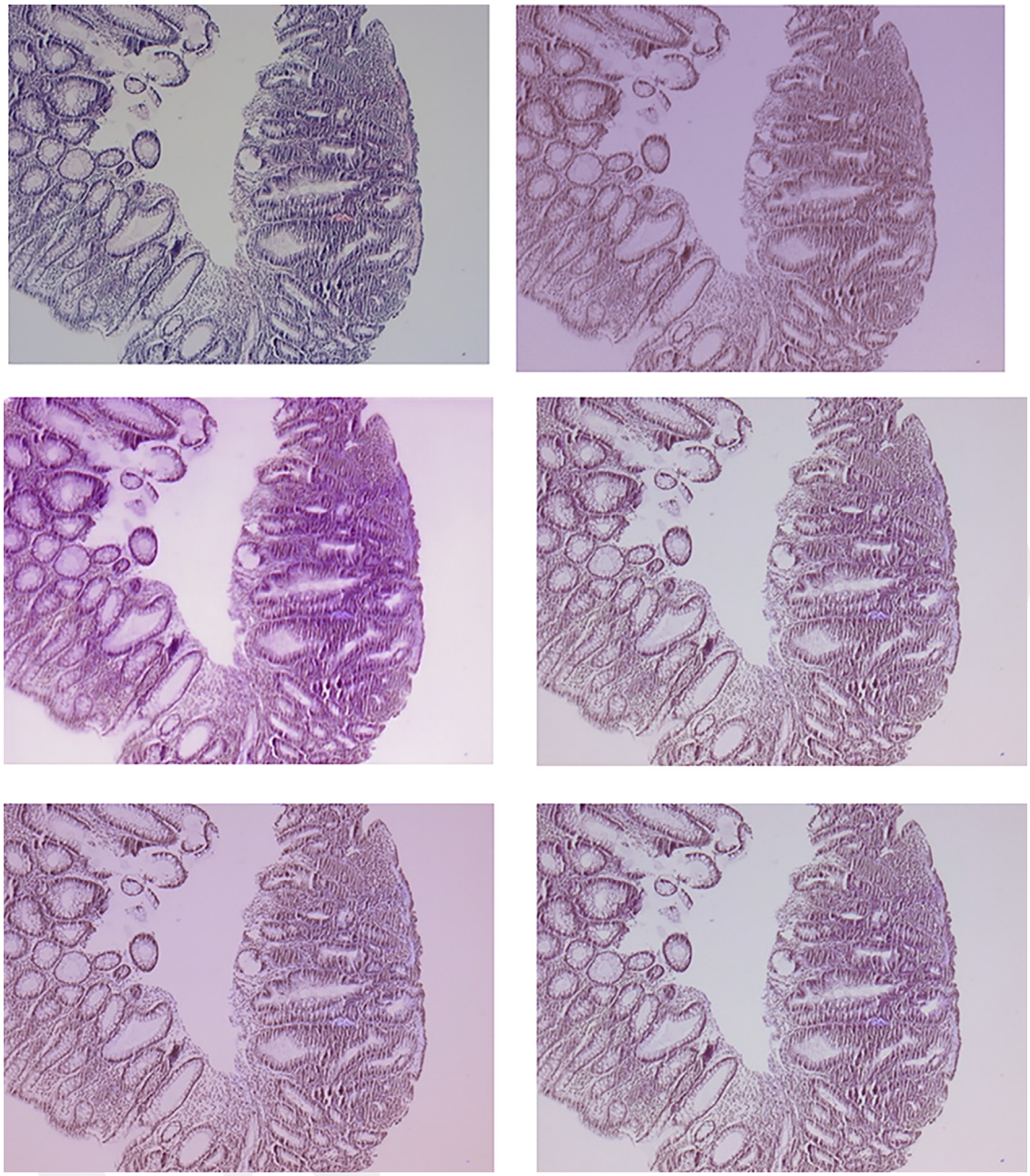

Fig. 2. Without any stain normalization, the color intensity variation between two different WSIs can be clearly seen in our custom dataset.

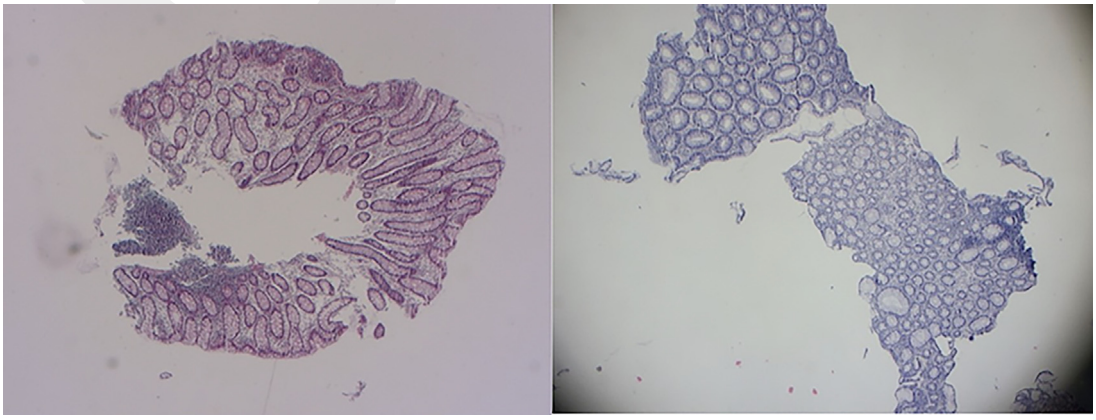

Fig. 3. Results of the stain normalization techniques on a sample image from our dataset.

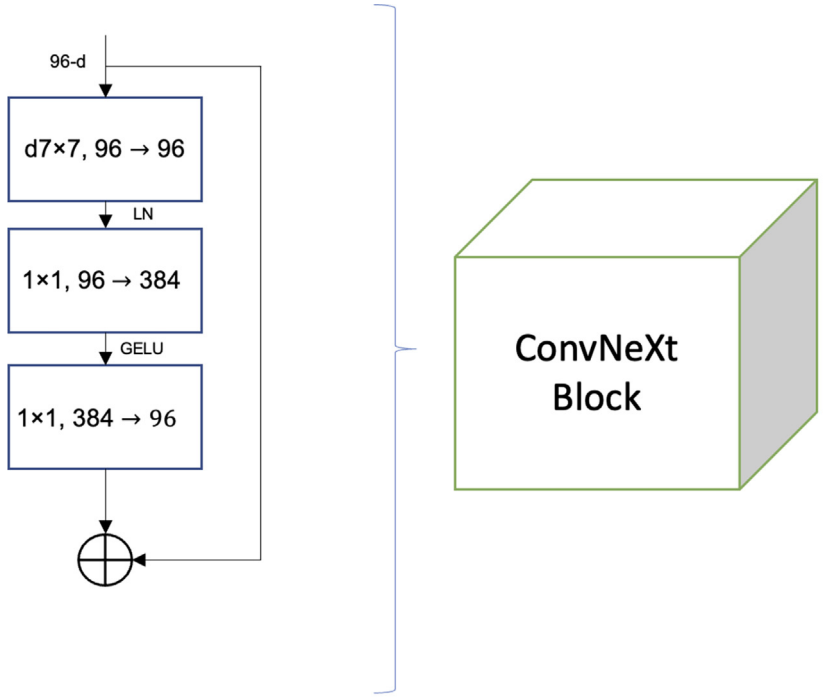

Fig. 4. Structure of a ConvNeXt block.

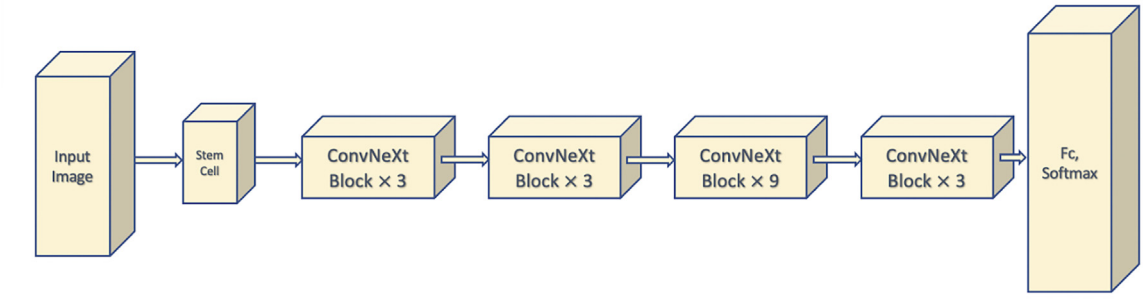

Fig. 5. The architecture of the ConvNeXt.

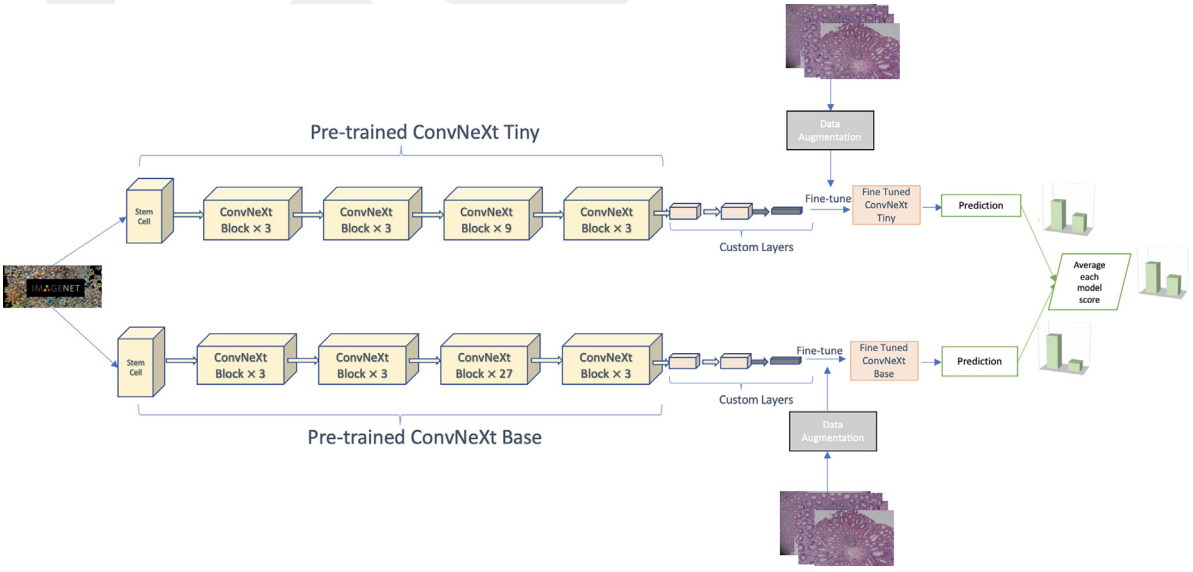

Fig. 6. The proposed framework of ensemble of ConvNeXts.

our task, we implemented a fine-tuning approach which consists of unfreezing the entire model and re-training it on our data for each of the models separately. Subsequently, drop-out and a dense layer are added to the top layer. Adam optimizer with a learning rate of 0.001 is employed, and adaptive momentum optimization algorithm optimized the learning rate during the training. As the loss function, binary cross entropy is used with label smoothing with a smoothing coefficient having a value of 0.1. By this way, label smoothing regularizes the network to choose the class with high confidence. The training batch size is set to 64 and both of the networks are fine-tuned in 50 epochs individually.

In the final decision step, the probabilistic outputs of each class, C_n, are calculated for base models, B_m. The average probabilistic output for each class is calculated at the averaging layer by using the outputs of the base models, as in Eq. 1.

$$P_{Ensemble}(C_i) = \frac{1}{|M|} \sum_{m \in M} P_{B_m}(C_i) \quad (1)$$

Furthermore, in order to verify the proposed model's generalizability, the model is tested on three different sets of instances, which come from UniToPatho, EBHI datasets and the testing set of the custom-collected dataset.

Additionally, in order to show the explainability of the proposed model, the Grad-CAM method is used. The attention map of the model is explored for adenomatous and non-adenomatous images. Outputs of the Grad-CAM results of the proposed model can help the pathologist to see and judge the inner decision step of the model.

3. Experimental setup

Fig. 6 shows the experimental setup of the proposed framework. The source code of the proposed algorithm and the experimental setup can be accessed online [47].

As it is mentioned in the Section 2.5; each of the base classifiers of the ensemble model is fine-tuned on our dataset by adding a drop-out and a dense layer to the top layer.

T performance of the proposed ensemble method is compared against the extensively used pre-trained deep CNN methods and attention-based method. We further compared frequently used stain normalization techniques for each of the deep CNN-based, attention-based methods and the proposed ensemble method. For these ablation tests the following standard performance measures are used: F1 score, accuracy, precision, and recall.

To prevent any "Clever Hans" type of mistake, the datasets used in the training and tests are split based on patients rather than individual images. That is, if a patient's sample image is in the test set, none of its other images can appear in the training set.

The models for the ablation study were employed from TensorFlow-Hub and the models are: Inception-V3 (InceptionV3 trained on ImageNet) ResNetV2-50 (ResNetV2-50 trained on ImageNet) ResNetV2-101, (ResNetV2-101 trained on ImageNet) InceptionResNet-V2 (InceptionResNet-V2 trained on ImageNet) ViT (fine-tuned on ImageNet 1k) EfficientNet-S (EfficientNet V2 pretrained on ImageNet), EfficientNet-S-21k (EfficientNet V2 pretrained on ImageNet 21k), EfficientNet-S-21k-ft-1k (EfficientNet V2 pretrained on ImageNet 21k and fine-tuned on ImageNet 1k) ConvNeXt-Tiny (model pre-trained on the ImageNet-1k dataset), ConvNeXt-Small (model pre-trained on the ImageNet-1k dataset), ConvNeXt-Base-1k (model pre-trained on the ImageNet-1k dataset), ConvNeXt-Base-21k (model pre-trained on the ImageNet-21k dataset), ConvNeXt-Base 21k-ft-1k (model pre-trained on ImageNet 21k and finetuned on ImageNet 1k), ConvNeXt-Large (model pre-trained on the ImageNet-21k dataset).

Table 3
Parameters of the proposed method.

Parameters	Values
Optimizer	ADAM
Learning Rate	0.001
Number of Epochs	50
Batch Size	64
Regularizer	L2 Norm

Table 4
Number of parameters of the models that are used in this work.

Network	Number of Parameters
ConvNeXt-Large	229,843,637
ConvNeXt-Base	87,568,514
ConvNeXt-Small	49,456,226
ConvNeXt-Tiny	27,821,666
Inception-v3	21,806,882
ResNet-v2-50	23,568,898
ResNet-v2-101	42,630,658
InceptionResNet-v2	54,339,810
ViT	36,047,682
EfficientNet-v2-s	20,333,922
Proposed Method	115,390,180

We employed all the models as pre-trained; this is due to the fact that building those models from scratch requires a large amount of data and resources.

After adding custom layers at the end of the base models, we implemented a fine-tuning approach because the state-of-the-art CNN models are pre-trained on natural images, while our images belong to a different domain. Thus, in order to comprehensively compare the above-mentioned methods, we first fine-tuned them on our histological dataset. All the experiments were performed on Google Colab Platform with 52 GB of RAM and NVIDIA Tesla K80, NVIDIA Tesla T4 and NVIDIA Tesla P100 GPU accelerators. The application of the proposed experiments was implemented with Python v3.7.13 with the TensorFlow v2.8.0 framework (See [47]).

The details of the network parameters, optimizers, number of epochs, learning rates, and number of parameters are given Table 3 and Table 4.

Initially, we first normalized the WSIs using the following stain normalization techniques separately and obtained different sets of normalized histological images: Stain-Net, Stain-GAN, Reinhard, Macenko and Vahadane. In order to apply Reinhard, Macenko and Vahadane techniques, we employed StainTools library [48]. The Stain-Net and Stain-GAN methods are implemented by using the source codes of [33], and [36].

4. Results

For the curated sets of WSIs with different stain normalization techniques, we primarily experiment with the aforementioned baseline classifiers. The ResNet-50 is implemented as it is proposed by Korbar et al. [16]. Moreover, other popular deep learning approaches are also implemented and the test performance of each model and stain normalization techniques are shown in Table 5 and Tables A.8–A.13. The performance results of the top-performed model on the custom dataset are given in Table 6. This table shows the accuracy metrics of the proposed model on the custom-collected dataset with various normalization techniques.

From Table 5 and Fig. 7 it can be seen that Stain-GAN and Reinhard normalization techniques perform better than other methods. Furthermore, accuracy and F1 scores of the Stain-GAN and Reinhard normalized datasets are approximately improved by 3–5% for

Table 5
Accuracy results of the baseline models and the proposed model on the curated sets of custom collected dataset.

	Original		Stain-Net		Stain-GAN		Reinhard		Macenko		Vahandane	
	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)
ConvNeXt-L	84.38	65.54	83.13	74.25	88.75	74.42	86.88	80.00	84.38	65.54	81.25	69.41
ConvNeXt-B-21k	80.63	73.94	86.25	85.00	80.00	58.89	86.67	75.71	74.38	49.73	57.50	37.36
ConvNeXt-B-21k-ft-1k	81.25	63.64	66.25	43.48	80.63	58.56	88.13	79.04	88.13	83.44	54.38	40.00
ConvNeXt-S	78.13	63.58	81.25	63.64	83.75	65.91	86.88	77.84	83.75	67.82	71.88	67.59
ConvNeXt-T	**87.50**	77.38	85.00	72.94	83.75	73.81	**91.36**	89.02	88.75	78.57	82.50	72.62
Inception-v3	82.50	68.60	80.00	68.24	85.53	69.77	86.25	74.12	82.50	66.67	75.74	63.44
ResNet-v2-50	76.25	60.92	78.13	63.58	82.50	64.77	82.50	64.77	75.63	54.14	77.50	62.07
ResNet-v2-101	79.81	65.52	77.36	56.98	83.13	72.19	80.00	58.89	74.21	53.93	71.25	49.45
InceptionResNet-v2	86.25	76.19	81.88	71.01	82.50	74.70	83.02	71.43	79.38	57.46	80.00	70.24
ViT	78.13	51.34	75.63	54.14	48.13	47.20	55.33	56.74	59.12	53.66	58.13	68.46
EfficientNet-v2-s	86.25	85.00	88.68	88.05	87.50	79.52	89.38	84.66	85.00	72.94	83.13	76.36
EfficientNet-v2-s-21k-ft-1k	85.00	83.75	86.88	82.21	**90.00**	86.42	89.87	82.93	**89.31**	87.50	83.75	81.08
EfficientNet-v2-s-21k	85.63	87.90	**89.38**	78.11	89.44	86.42	89.38	78.11	88.75	87.50	84.13	71.62
Proposed Method	**93.75**	93.58	**95.00**	93.90	**92.50**	91.25	**91.88**	90.57	**91.93**	90.48	**88.82**	87.12

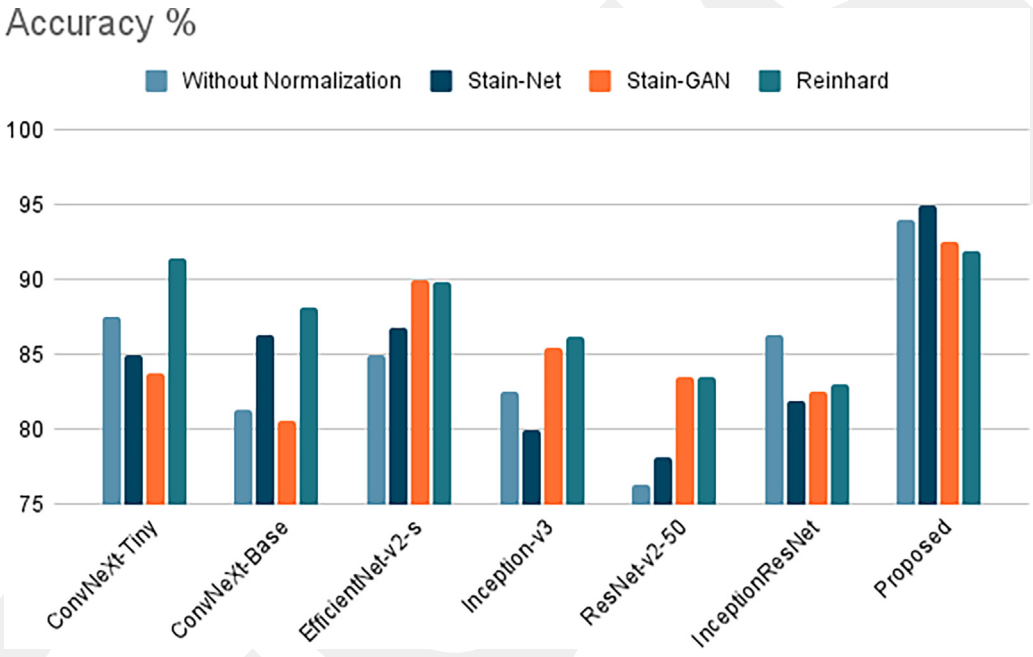

Fig. 7. Comparison of different stain normalization techniques for the proposed method and baseline models.

Table 6
Performance results of the proposed model.

	Proposed Method			
Normalization	Accuracy	Precision	Recall	F1
Original	93.75%	93.58%	93.58%	93.58%
Stain-Net	95.00%	92.77%	95.06%	93.90%
Stain-Gan	92.50%	92.41%	90.12%	91.25%
Reinhard	91.88%	92.31%	88.89%	90.57%
Vahandane	88.82%	87.65%	86.59%	87.12%
Macenko	91.93%	88.37%	92.68%	90.48%

the baseline models. The performance of the proposed model is evaluated for each stain normalization technique and is given in Table 5.

The proposed method on our custom dataset performs the best on Stain-Net normalized dataset and achieves the highest accuracy, precision, recall, and F-score with values of 95%, 92.8%, 95.1% and 93.9%, respectively. On the other hand, the performance of the ensemble model is relatively poor for the Vahandane normalized dataset with accuracy, precision, recall, and F-score with values of 88.8%, 87.7%, 86.6% and 87.1%, respectively.

The performance of the proposed method and all the baseline classifiers are poor for the Vahandane normalized data. For the Vahandane normalized dataset, the same proportion of adenomatous and non-adenomatous images are confused by all the base classifiers and the proposed model. This may originate from the Vahandane normalized images having less contrast than the other normalized images, as shown in Fig. 3 and Fig. 3d. In order to address this problem, we implemented different data augmentation techniques that include random contrast, random brightness and random hue, however, we observe that it is more suitable to employ image pre-processing techniques, such as adaptive histogram equalization, before the normalization of an image with Vahandane normalization algorithm.

A comprehensive comparison of the top performed state-of-the-art Deep CNN classifiers' performance on the non-normalized,

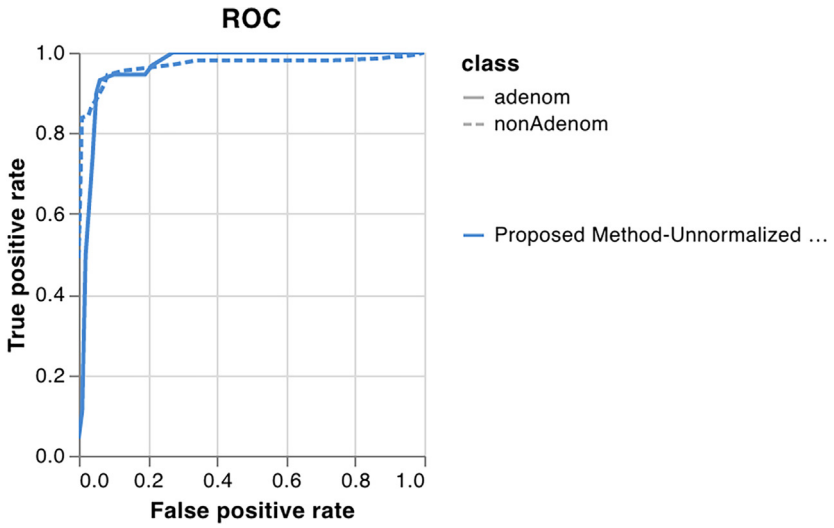

Fig. 8. ROC of the proposed method on our dataset without normalization.

Stain-GAN Normalized, Stain-Net normalized and Reinhard normalized data are presented in Fig. 7. As it can be seen from the figure and the table, for the Reinhard normalized dataset, the maximum accuracies of 91.4%, 88.1%, 89.9%, 86.2%, 83.5% and 83.0% are produced by ConvNeXt-Tiny, ConvNeXt-Base, Efficient-Net-v2-S, Inception-v3, ResNet-v2-50 and InceptionResNet models, respectively. Moreover, we can see that the performance of the single deep CNN classifiers' accuracy results varies in terms of different normalization techniques. However, the variation of performance for the proposed method for different normalization techniques is relatively small. Thus, this minor variation shows that the proposed model is more robust to input variations in the given datasets and generalizes better than the single model classifiers.

As it can be seen in Tables A.8 to A.13, the most significant performance gap for the proposed method and the base classifiers is obtained for the non-normalized and Stain-Net normalized data. The proposed method increased the overall accuracy for the non-normalized and Stain-Net normalized data by 6%, for Vahandane normalized data by 4%, and for the Stain-GAN and Macenko normalized data by 2%. The best accuracy of the ensemble method is achieved for the Stain-Net normalized data with an accuracy of 95% which is followed by Efficient-Net-v2-S with 89%. The ROC curves of the proposed method for the non-normalized dataset and Stain-Net normalized dataset are presented on Fig. 8 and Fig. 9, respectively. The results of the experiments show that the performance of the proposed ensemble model is satisfactory on all the normalized datasets and non-normalized dataset. Especially, the results on the non-normalized dataset show that the proposed ensemble model has a sufficient generalization ability since the images on the dataset differ in terms of color intensity.

4.1. Generalization test

In order to test the generalization ability of the model, colonic adenomatous and non-adenomatous images are employed from both the UniToPatho and EBHI datasets. The obtained images are fed into the model that is trained on our non-normalized dataset, with fine-tuning. Overall accuracies of 91.1% and 90% are achieved for EBHI and UniToPatho datasets, respectively. The performance metrics of the proposed model on UnitoPatho and EBHI are present in Table 7. ROC curves of the generalization test are given in Figs. 10, 11.

Table 7
Performance results of the proposed model on different datasets.

	Proposed Method			
Dataset	Accuracy	Precision	Recall	F1
Custom	95.00%	92.77%	95.06%	93.90%
UniToPatho	90.00%	91.83%	89.10%	90.45%
EBHI	91.1%	88.74%	94.36%	91.46%

Additionally, confusion matrices for the proposed method on UnitoPatho, EBHI and Custom Collected datasets are given in the Figs. 12, 13, and 14.

4.2. Grad-CAM results of the proposed method

To see the proposed models' class activations maps, gradient-weighted class activation mapping (Grad-CAM) method is employed [49]. Grad-CAM method explains the operation of a deep model by using the activation maps of a model in which the more focused regions are highlighted with a red color while the less attention grasping regions highlighted with yellow to blue colors. The Figs. 15 and 16 show the Grad-CAM results of the proposed model, which ensembles ConvNeXt-Tiny and ConvNeXt-Base. As it can be seen from the figures, the proposed model focuses on the spatially distant cell structures to classify an image. Furthermore, Grad-CAM outputs of the proposed model for the pathological image can provide insight to a pathologist by explaining why an image is classified as adenomatous or non-adenomatous by the model.

5. Discussion

In this work, we propose an ensemble method which employs the recently proposed ConvNeXt variants for polyp classification on Stain-Net normalized histopathology images. The proposed method combines two separately fine-tuned ConvNeXt variants, namely ConvNeXt-Tiny and ConvNeXt-Base. The base models are tailored to the classification problem by network modifications at the image representation levels. The performance of the ensemble method is compared with the state-of-the-art deep CNN mod-

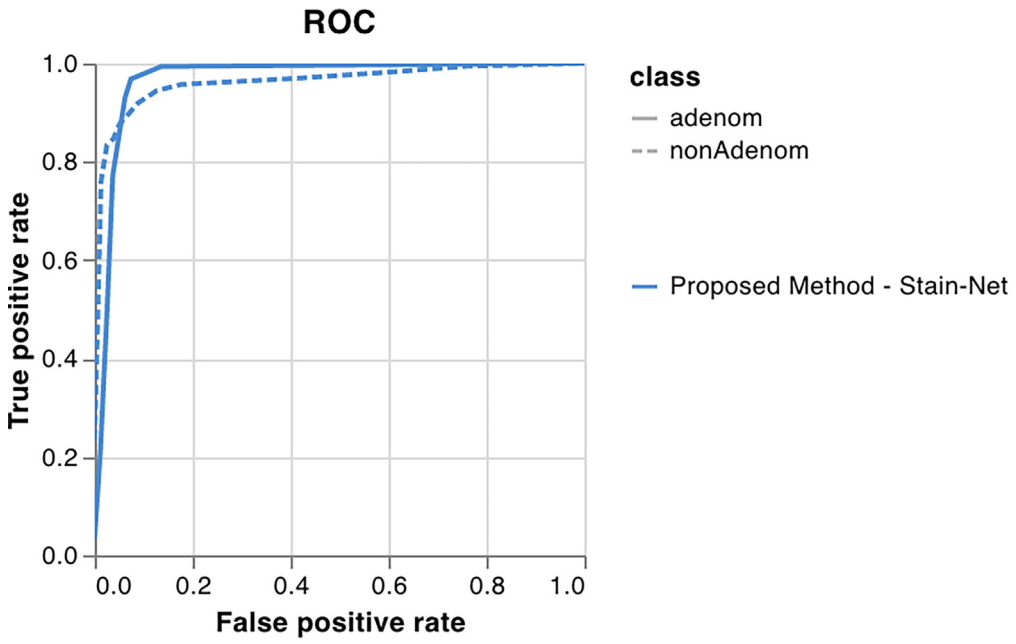

Fig. 9. ROC of the proposed method on our dataset with Stain-Net normalization.

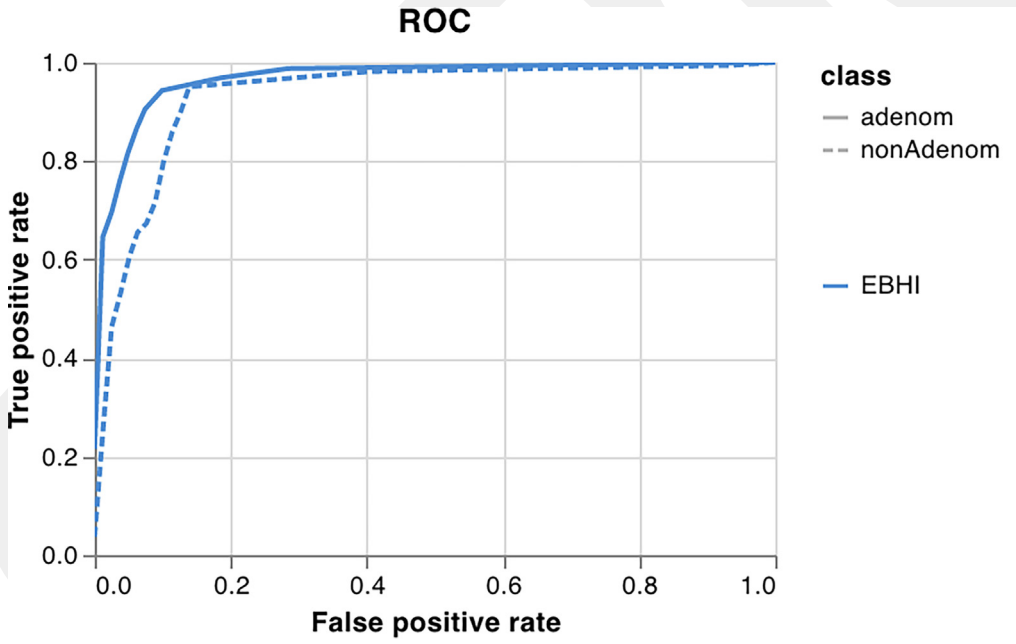

Fig. 10. ROC of the proposed method on EBHI dataset.

els and attention-based models on a custom colonic histological dataset. As a result, comprehensive experiments indicate that the ensemble of baseline models performs better than baseline models alone.

Ensemble methods are used in cancer classification tasks on breast, and colon histology images [7,14,26–29]. However, ensemble methods were not used in the colonic adenomatous polyp detection from the histology images. In the literature, researchers generally employ Deep CNN models alone. For example, in their work, Korbar et al. [16] employed various ResNet-50 variants and selected the best-performed variant, which achieved 91.3% accuracy on their dataset. Byeon et al. implemented EfficientNet for colon polyp subtype classification and achieved an overall F1 score of 98.8 on their dataset [50]. Iizuka et al. employed Inception-V3 to differentiate adenomatous, non-adenomatous and cancerous tissues on histopathology images and achieved an accuracy of 96% [6].

During the experimental setup, we implemented ResNet50, EfficientNet, Inception-v3 and compared the performance with our proposed method on the custom collected dataset. The proposed method achieves an accuracy of 93.75%, while ResNet50, EfficientNet and Inception-v3 achieve accuracies of 76.25%, 86.25% and

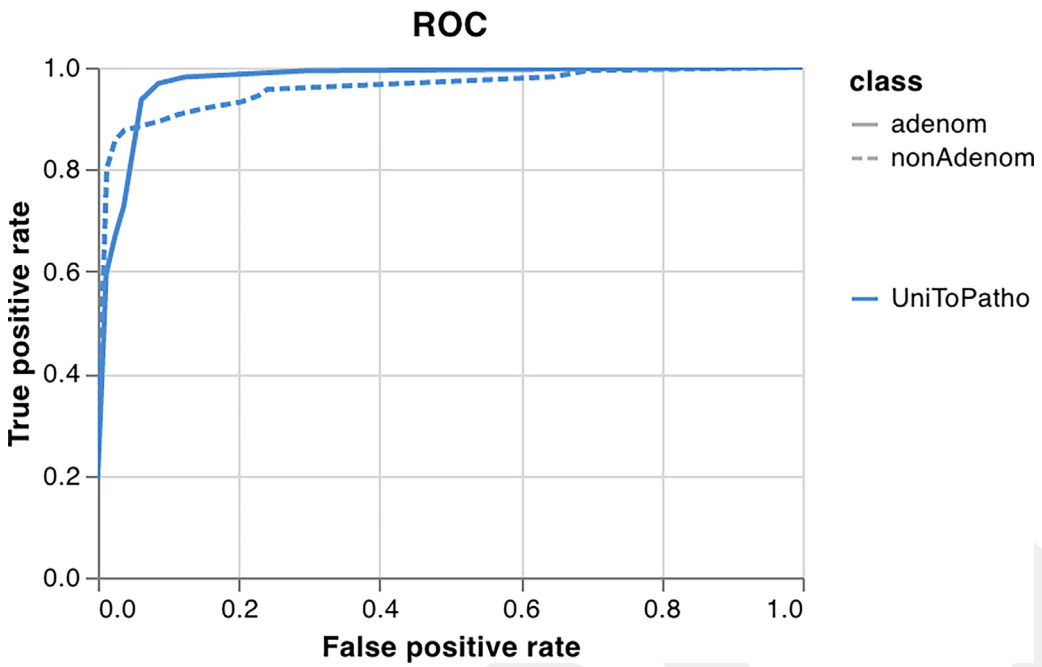

Fig. 11. ROC of the proposed method on UniToPatho dataset.

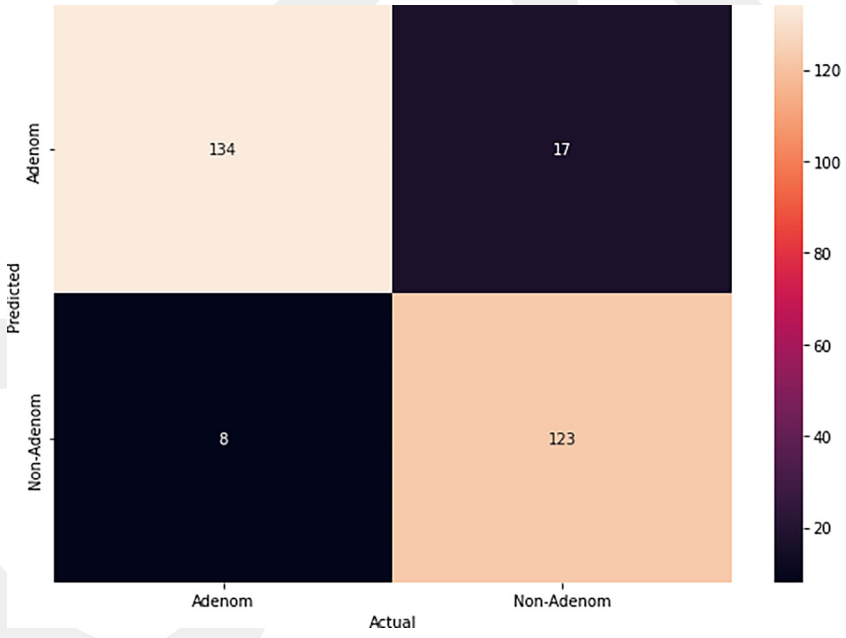

Fig. 12. Confusion Matrix of the proposed method on EBHI dataset.

82.5% on the custom collected dataset, respectively. The gap in the model's performance on our custom-collected dataset and their dataset may be caused by the different domain distributions of the datasets.

In the literature, the models are generally tested against specific custom datasets. Since the models are developed for a specific dataset, they may not work well on the other datasets. Thus, this might be a drawback for real-world applications. To overcome this issue, researchers use publicly available datasets for benchmarking the models that are built for a custom dataset [2,7,22,23,51]. Therefore, in this work, additional experiments are conducted to explore the performance of the proposed model on two publicly available datasets, UniToPatho and EBHI. The proposed method outperforms the other methods by attaining 90% and 91.1% on UniToPatho and EBHI, while other methods in the literature achieve an accuracy of 64.29% and 66.55% on UniToPatho dataset [22,23]. These accuracy results demonstrate that the proposed model has promising generalization ability for different datasets and has the potential to work in real-life scenarios.

To increase the models' generalization ability, stain normalization techniques are widely employed by researchers on HI. In contrast to the previous studies which make polyp classifica-

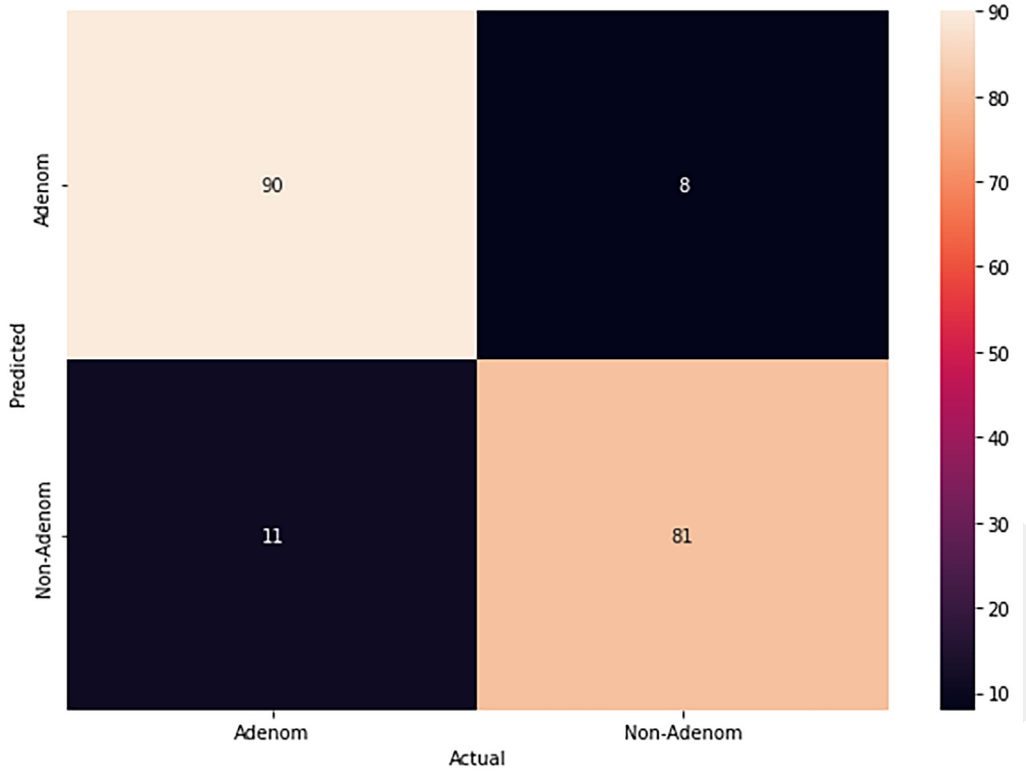

Fig. 13. Confusion Matrix of the proposed method on UniToPatho dataset.

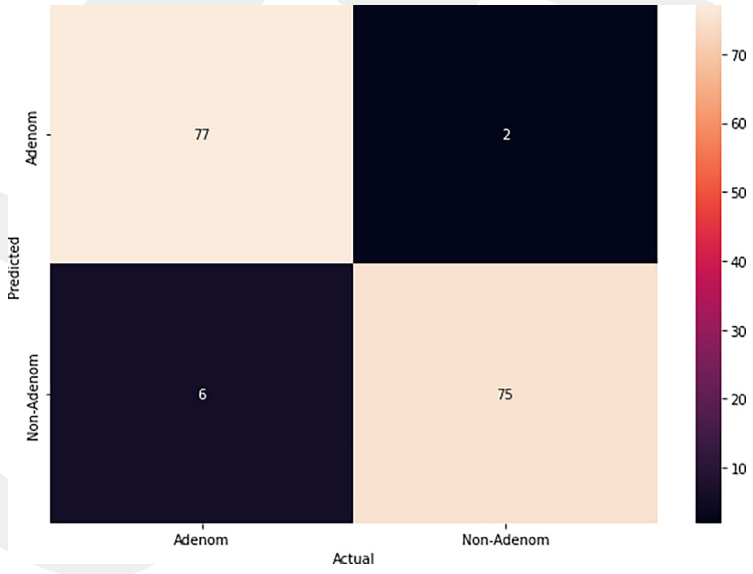

Fig. 14. Confusion Matrix of the proposed method on Custom Collected dataset.

tion on HI, in this study, stain normalization techniques are combined with an ensemble model. To the best of our knowledge, there is a limited number of studies which incorporates stain normalization methods for colon polyp classification on histopathology images [20]. In the literature, Perlo et al. use only Macenko normalization technique for polyp classification on histopathology images [20]. However, during the experiments, we observed that the performance of normalization techniques significantly differs for different classifiers. Therefore, combining various classifiers with different stain normalization techniques produced more beneficial outputs.

When it comes to medical image analysis, the black-box nature of the AI methods might restrict their usage in real applications. In recent years, this has sparked debates about the usage and necessity of explainability of opaque algorithms [52]. There have been efforts to overcome this problem by introducing several tools for

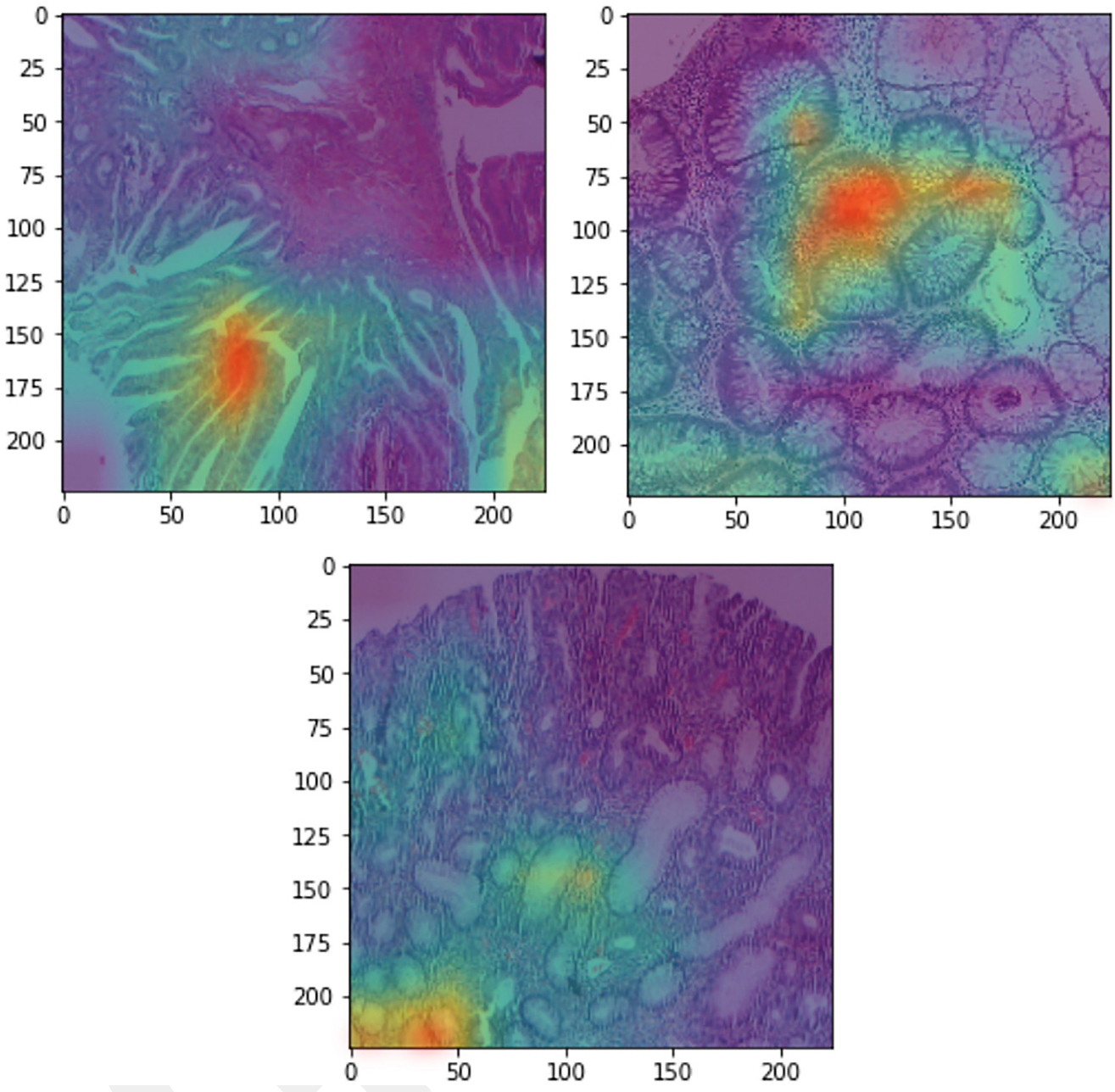

Fig. 15. Grad-CAM results of adenomatous images.

contemporary deep learning models [53]. The main approach to solve the problem is to provide the underlying reason for the decision as an auxiliary output to the clinician. This output could be either verbal or visual cues. This would be useful, especially when there is a mismatch between the clinician's and the system's decisions. In this case, the visual output might help to resolve the conflict. The clinician can check why the system diagnosed differently by evaluating the cues about the decision process of the system. As an interpretability method, the proposed system highlights the image regions which affected the decision most. As it was used in other medical image computing applications [54], an attribution-based explainability method, Grad-Cam, is employed. Various studies use Grad-Cam to assists during the decision-making process of pathologists [2,6,11,15,17,21,50,55,56]. In their work, [21] passed the Grad-Cam outputs to the expert pathologists to evaluate the models' performance to find a model that approximates most to the human interpreters [2]. provided Grad-Cam outputs to experts for evaluation of their model. Further, [56] and [55] used Grad-Cam method to provide the explainability of the model. In their work, [6] used Grad-Cam outputs to compare their models' performance with pathologists and medical school students [50]. evaluated their model by using the Grad-Cam outputs for different polyp types [15]. annotated the region of interest using Grad-Cam [17]. provided Grad-Cam results to pathologists for decision support.

In this study, the interpretability of the Grad-Cam outputs is found to be mostly in line with the pathologist's expectations in the initial tests. However, the full investigation and an objective evaluation of the explainability of the system on the full dataset are left for future study.

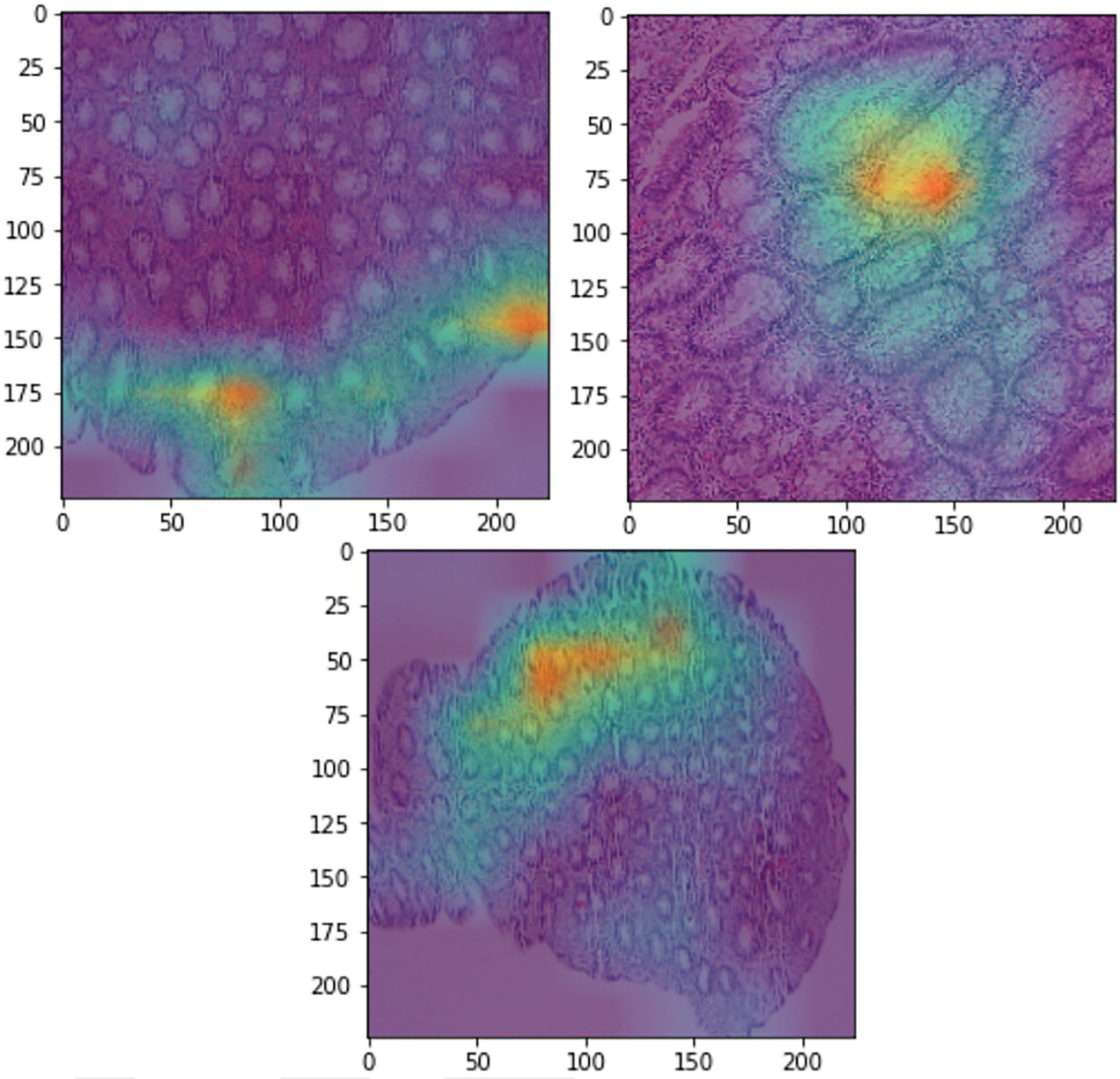

Fig. 16. Grad-CAM results of non-adenomatous images.

In future work, by collecting more colonic histological images, the performance of the proposed method can be explored for multi-class classification of the adenomatous images. Moreover, for the multi-class classification purpose outputs of the base classifiers can be combined in a weighted fashion, so that the ensemble model can give more weight to a better-performing base classifier for a specific task.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Sena Busra Yengec Tasdemir reports financial support was provided by Scientific and Technological Research Council of Turkey.

CRediT authorship contribution statement

Sena Busra Yengec-Tasdemir: Data curation, Conceptualization, Methodology, Writing – original draft. **Zafer Aydin:** Writing – review & editing. **Ebru Akay:** Data curation. **Serkan Dogan:** Data curation. **Bulent Yilmaz:** Supervision, Writing – review & editing.

Acknowledgments

This work was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under Grant 120E204.

The authors would like to thank Serdal Sadet Ozcan for her valuable contribution to detailed labeling of the dataset.

Appendix A. Tables

Table A.8
Accuracy results of the baseline models and the proposed model on our dataset without normalization.

Classifier	Accuracy	Precision	Recall	F1 Score
ConvNeXt-L	84.38%	73.42%	59.18%	65.54%
ConvNeXt-B-21k	80.63%	77.22%	70.93%	73.94%
ConvNeXt-B-21k-ft-1k	81.25%	70.89%	57.73%	63.64%
ConvNeXt-S	78.13%	69.62%	58.51%	63.58%
ConvNeXt-T	87.50%	82.28%	73.03%	77.38%
Inception-v3	82.50%	74.68%	63.44%	68.60%
ResNet-v2-50	76.25%	67.09%	55.79%	60.92%
ResNet-v2-101	79.81%	69.72%	61.79%	65.52%
InceptionResNet-v2	86.25%	81.01%	71.91%	76.19%
ViT	78.13%	60.76%	44.44%	51.34%
EfficientNet-v2-s	86.25%	86.08%	83.95%	85.00%
EfficientNet-v2-s-21k-ft-1k	85.00%	84.81%	82.72%	83.75%
EfficientNet-v2-s-21k	85.63%	87.34%	88.46%	87.90%
Proposed Method	93.75%	93.58%	93.58%	93.58%

Table A.9
Accuracy results of the baseline models and the proposed model on our dataset with Stain-Net normalization.

Classifier	Accuracy	Precision	Recall	F1 Score
ConvNeXt-L	83.13%	78.48%	70.45%	74.25%
ConvNeXt-B-21k-ft-1k	86.25%	86.08%	83.95%	85.00%
ConvNeXt-B-21k-ft-1k	66.25%	50.63%	38.10%	43.48%
ConvNeXt-S	81.25%	70.89%	57.73%	63.64%
ConvNeXt-T	85.00%	78.48%	68.13%	72.94%
Inception-v3	80.00%	73.42%	63.74%	68.24%
ResNet-v2-50	78.13%	69.62%	58.51%	63.58%
ResNet-v2-101	77.36%	64.56%	51.00%	56.98%
InceptionResNet-v2	81.88%	75.95%	66.67%	71.01%
ViT	75.63%	62.03%	48.04%	54.14%
EfficientNet-v2-s	88.68%	88.61%	87.50%	88.05%
EfficientNet-v2-s-21k-ft-1k	86.88%	84.81%	79.76%	82.21%
EfficientNet-v2-s-21k	89.38%	83.54%	73.33%	78.11%
Proposed Method	95.00%	92.77%	95.06%	93.90%

Table A.10
Accuracy results of the baseline models and the proposed model on our dataset with Stain-GAN normalization.

Classifier	Accuracy	Precision	Recall	F1 Score
ConvNeXt-L	88.75%	81.01%	68.82%	74.42%
ConvNeXt-B-21k-ft-1k	80.00%	67.09%	52.48%	58.89%
ConvNeXt-B-21k-ft-1k	80.63%	67.09%	51.96%	58.56%
ConvNeXt-S	83.75%	73.42%	59.79%	65.91%
ConvNeXt-T	83.75%	78.48%	69.66%	73.81%
Inception-v3	85.53%	76.92%	63.83%	69.77%
ResNet-v2-50	82.50%	72.15%	58.76%	64.77%
ResNet-v2-101	83.13%	77.22%	67.78%	72.19%
InceptionResNet-v2	82.50%	78.48%	71.26%	74.70%
ViT	48.13%	47.50%	46.91%	47.20%
EfficientNet-v2-s	87.50%	83.54%	75.86%	79.52%
EfficientNet-v2-s-21k-ft-1k	90.00%	88.61%	84.34%	86.42%
EfficientNet-v2-s-21k	89.44%	88.61%	84.34%	86.42%
Proposed Method	92.50%	92.41%	90.12%	91.25%

Table A.11
Accuracy results of the baseline models and the proposed model on our dataset with Reinhard normalization.

Classifier	Accuracy	Precision	Recall	F1 Score
ConvNeXt-L	86.88%	83.54%	76.74%	80.00%
ConvNeXt-B-21k-ft-1k	86.67%	79.76%	72.04%	75.71%
ConvNeXt-B-21k-ft-1k	88.13%	83.54%	75.00%	79.04%
ConvNeXt-S	86.88%	82.28%	73.86%	77.84%
ConvNeXt-T	91.36%	90.12%	87.95%	89.02%
Inception-v3	86.25%	79.75%	69.23%	74.12%
ResNet-v2-50	82.50%	72.15%	58.76%	64.77%
ResNet-v2-101	80.00%	67.09%	52.48%	58.89%
InceptionResNet-v2	83.02%	76.92%	66.67%	71.43%
ViT	55.33%	57.97%	55.56%	56.74%
EfficientNet-v2-s	89.38%	87.34%	82.14%	84.66%
EfficientNet-v2-s-21k-ft-1k	89.87%	86.08%	80.00%	82.93%
EfficientNet-v2-s-21k	89.38%	83.54%	73.33%	78.11%
Proposed Method	91.88%	92.31%	88.89%	90.57%

Table A.12
Accuracy results of the baseline models and the proposed model on our dataset with Vahandane normalization.

Classifier	Accuracy	Precision	Recall	F1 Score
ConvNeXt-L	81.25%	74.68%	64.84%	69.41%
ConvNeXt-B-21k-ft-1k	57.50%	43.04%	33.01%	37.36%
ConvNeXt-B-21k-ft-1k	54.38%	44.30%	36.46%	40.00%
ConvNeXt-S	71.88%	76.56%	60.49%	67.59%
ConvNeXt-T	82.50%	77.22%	68.54%	72.62%
Inception-v3	75.74%	67.05%	60.20%	63.44%
ResNet-v2-50	77.50%	68.35%	56.84%	62.07%
ResNet-v2-101	71.25%	56.96%	43.69%	49.45%
InceptionResNet-v2	80.00%	74.68%	66.29%	70.24%
ViT	58.13%	64.56%	72.86%	68.46%
EfficientNet-v2-s	83.13%	79.75%	73.26%	76.36%
EfficientNet-v2-s-21k-ft-1k	83.75%	89.55%	74.07%	81.08%
EfficientNet-v2-s-21k	84.13%	75.23%	68.33%	71.62%
Proposed Method	88.82%	87.65%	86.59%	87.12%

Table A.13
Accuracy results of the baseline models and the proposed model on our dataset with Macenko normalization.

Classifier	Accuracy	Precision	Recall	F1 Score
ConvNeXt-L	84.38%	73.42%	59.18%	65.54%
ConvNeXt-B-21k-ft-1k	74.38%	58.23%	43.40%	49.73%
ConvNeXt-B-21k-ft-1k	88.13%	86.08%	80.95%	83.44%
ConvNeXt-S	83.75%	74.68%	62.11%	67.82%
ConvNeXt-T	88.75%	83.54%	74.16%	78.57%
Inception-v3	82.50%	73.42%	61.05%	66.67%
ResNet-v2-50	75.63%	62.03%	48.04%	54.14%
ResNet-v2-101	74.21%	61.54%	48.00%	53.93%
InceptionResNet-v2	79.38%	65.82%	50.98%	57.46%
ViT	59.12%	55.70%	51.76%	53.66%
EfficientNet-v2-s	85.00%	78.48%	68.13%	72.94%
EfficientNet-v2-s-21k-ft-1k	89.31%	88.61%	86.42%	87.50%
EfficientNet-v2-s-21k	88.75%	88.61%	86.42%	87.50%
Proposed Method	91.93%	88.37%	92.68%	90.48%

References

[1] H. Sung, J. Ferlay, R.L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, et al., Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA: Cancer J. Clin. 71 (3) (2021) 209–249, doi:10.3322/CAAC.21660.

[2] M. Bilal, Y.W. Tsang, M. Ali, S. Graham, E. Hero, N. Wahab, K. Dodd, H. Sahota, W. Lu, M. Jahanifar, A. Robinson, A. Azam, K. Benes, M. Nimir, A. Bhalerao, H. Eldaly, S.E.A. Raza, K. Gopalakrishnan, F. Minhas, D. Snead, N. Rajpoot, AI Based pre-screening of large bowel cancer via weakly supervised learning of colorectal biopsy histology images, medRxiv (2022), doi:10.1101/2022.02.28.22271565.

[3] D. Bychkov, N. Linder, R. Turkki, S. Nordling, P.E. Kovanen, C. Verrill, M. Walliander, M. Lundin, C. Haglund, J. Lundin, et al., Deep learning based tissue analysis predicts outcome in colorectal cancer, Sci. Rep. 2018 8:1 8 (1) (2018) 1–11, doi:10.1038/s41598-018-21758-3.

[4] P. Gupta, Y. Huang, P.K. Sahoo, J.F. You, S.F. Chiang, D.D. Onthoni, Y.J. Chern, K.Y. Chao, J.M. Chiang, C.Y. Yeh, W.S. Tsai, Colon tissues classification and localization in whole slide images using deep learning, Diagnostics (Basel) 11 (8) (2021), doi:10.3390/DIAGNOSTICS11081398.

[5] C. Ho, Z. Zhao, X.F. Chen, J. Sauer, S.A. Saraf, R. Jialdasani, K. Taghipour, A. Sathe, L.Y. Khor, K.H. Lim, W.Q. Leow, et al., A promising deep learning-assistive algorithm for histopathological screening of colorectal cancer, Sci. Rep. 2022 12:1 12 (1) (2022) 1–9, doi:10.1038/s41598-022-06264-x.

[6] O. Iizuka, F. Kanavati, K. Kato, M. Rambeau, K. Arihiro, M. Tsuneki, Deep Learning Models for Histopathological Classification of Gastric and Colonic Epithelial Tumours, 2020. doi:10.1038/s41598-020-58467-9.

[7] A. Kallipolitis, K. Revelos, I. Maglogiannis, Ensembling efficientnets for the classification and interpretation of histopathology images, Algorithms 14 (10) (2021), doi:10.3390/a14100278.

[8] D. Sarwinda, R.H. Paradisa, A. Bustamam, P. Anggia, et al., Deep Learning in Image Classification using Residual Network (ResNet) Variants for Detection of Colorectal Cancer, in: Procedia Computer Science, volume 179, Elsevier, 2021, pp. 423–431, doi:10.1016/j.procs.2021.01.025.

[9] T.E. Tavolara, M.K.K. Niazi, V. Arole, W. Chen, W. Frankel, M.N. Gurcan, A modular cGAN classification framework: application to colorectal tumor detection, Sci. Rep. 9 (1) (2019) 1–8, doi:10.1038/s41598-019-55257-w.

[10] E. Terradillos, C.L. Saratxaga, S. Mattana, R. Cicchi, F.S. Pavone, N. Andraka, B.J. Glover, N. Arbide, J. Velasco, M.C. Etxezarraga, A. Picon, Analysis on the characterization of multiphoton microscopy images for malignant neoplastic colon lesion detection under deep learning methods, J. Pathol. Inform. 12 (1) (2021) 27, doi:10.4103/jpi.jpi_113_20.

[11] M. Tsuneki, F. Kanavati, diagnostics Deep Learning Models for Poorly Differentiated Colorectal Adenocarcinoma Classification in Whole Slide Images Using Transfer Learning (2021). doi:10.3390/diagnostics11112074.

[12] M. Yildirim, A. Cinar, Classification with respect to colon adenocarcinoma and colon benign tissue of colon histopathological images with a new CNN MA_colonNET, Int. J. Imaging Syst. Technol. 32 (1) (2022) 155–162, doi:10.1002/ima.22623.

[13] C. Zhou, Y. Jin, Y. Chen, S. Huang, R. Huang, Y. Wang, Y. Zhao, Y. Chen, L. Guo, J. Liao, Histopathology classification and localization of colorectal cancer using global labels by weakly supervised deep learning, Comput. Med. Imag. Graphic. 88 (2021) 101861, doi:10.1016/J.COMPMEDIMAG.2021.101861.

[14] D. Albashish, Ensemble of adapted convolutional neural networks (CNN) methods for classifying colon histopathological images, PeerJ Comput. Sci. 8 (2022) e1031.

[15] B. Korbar, A.M. Olofson, A.P. Miraflor, C.M. Nicka, M.A. Suriawinata, L. Torresani, A.A. Suriawinata, S. Hassanpour, Looking under the hood: deep neural network visualization to interpret whole-Slide image analysis outcomes for colorectal polyps, IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops 2017-July (2017) 821–827, doi:10.1109/CVPRW.2017.114.

[16] B. Korbar, A. Olofson, A. Miraflor, C. Nicka, M. Suriawinata, L. Torresani, A. Suriawinata, S. Hassanpour, et al., Deep learning for classification of colorectal polyps on whole-slide images, J. Pathol. Inform. 8 (1) (2017), doi:10.4103/JPI.JPI_34_17.

[17] Z. Song, C. Yu, S. Zou, W. Wang, Y. Huang, X. Ding, J. Liu, L. Shao, J. Yuan, X. Gou, W. Jin, Z. Wang, X. Chen, H. Chen, C. Liu, G. Xu, Z. Sun, C. Ku, Y. Zhang, X. Dong, S. Wang, W. Xu, N. Lv, H. Shi, Automatic deep learning-based colorectal adenoma detection system and its similarities with pathologists, BMJ Open 10 (9) (2020) e036423, doi:10.1136/BMJOPEN-2019-036423.

[18] J.W. Wei, A.A. Suriawinata, L.J. Vaickus, B. Ren, X. Liu, M. Lisovsky, N. Tomita, B. Abdollahi, A.S. Kim, D.C. Snover, J.A. Baron, E.L. Barry, S. Hassanpour, et al., Evaluation of a deep neural network for automated classification of colorectal polyps on histopathologic slides, JAMA Netw. Open 3 (4) (2020), doi:10.1001/JAMANETWORKOPEN.2020.3398.

[19] M. Nasir-Moin, A.A. Suriawinata, B. Ren, X. Liu, D.J. Robertson, S. Bagchi, N. Tomita, J.W. Wei, T.A. Mackenzie, J.R. Rees, S. Hassanpour, Evaluation of an artificial intelligence-Augmented digital system for histologic classification of colorectal polyps, JAMA Netw. Open 4 (11) (2021) 1–12, doi:10.1001/jamanetworkopen.2021.35271.

[20] D. Perlo, E. Tartaglione, L. Bertero, P. Cassoni, M. Grangetto, Dysplasia grading of colorectal polyps through CNN analysis of WSI (2021). doi:10.48550/arxiv.2102.05498.

[21] J. Wei, A. Suriawinata, B. Ren, X. Liu, M. Lisovsky, L. Vaickus, C. Brown, M. Baker, M. Nasir-Moin, N. Tomita, L. Torresani, J. Wei, S. Hassanpour, et al., Learn like a Pathologist: Curriculum Learning by Annotator Agreement for Histopathology Image Classification(2021).

[22] X. Wang, S. Yang, J. Zhang, M. Wang, J. Zhang, W. Yang, J. Huang, X. Han, Transformer-based unsupervised contrastive learning for histopathological image classification, Med. Image Anal. 81 (2022) 102559.

[23] X. Wang, Y. Du, S. Yang, J. Zhang, M. Wang, J. Zhang, W. Yang, J. Huang, X. Han, RetCCL: clustering-guided contrastive learning for whole-slide image retrieval, Med. Image Anal. 83 (2023) 102645.

[24] W.S. Liew, T.B. Tang, C.-H. Lin, C.-K. Lu, Automatic colonic polyp detection using integration of modified deep residual convolutional neural network and ensemble learning approaches, Comput. Methods Programs Biomed. 206 (2021) 106114, doi:10.1016/j.cmpb.2021.106114.

[25] F. Younas, M. Usman, W.Q. Yan, A deep ensemble learning method for colorectal polyp classification with optimized network parameters, Appl. Intell. (2022) 1–24.

[26] S. Ghosh, A. Bandyopadhyay, S. Sahay, R. Ghosh, I. Kundu, K.C. Santosh, Colorectal histology tumor detection using ensemble deep neural network, Eng. Appl. Artif. Intell. 100 (2021) 104202, doi:10.1016/j.engappai.2021.104202.

[27] S. Mehmood, T.M. Ghazal, M.A. Khan, M. Zubair, M.T. Naseem, T. Faiz, M. Ahmad, Malignancy detection in lung and colon histopathology images using transfer learning with class selective image processing, IEEE Access 10 (2022) 25657–25668, doi:10.1109/ACCESS.2022.3150924.

[28] D.S. Luz, T.J.B. Lima, R.R.V. Silva, D.M.V. Magalhães, F.H.D. Araujo, Automatic detection metastasis in breast histopathological images based on ensemble learning and color adjustment, Biomed. Signal Process. Control 75 (2022) 103564, doi:10.1016/j.bspc.2022.103564.

[29] A. Kumar, J. Kim, D. Lyndon, M. Fulham, D. Feng, et al., An ensemble of fine-Tuned convolutional neural networks for medical image classification, IEEE J. Biomed. Health Inform. 21 (1) (2017) 31–40, doi:10.1109/JBHI.2016.2635663.

[30] C.A. Barbano, D. Perlo, E. Tartaglione, A. Fiandrotti, L. Bertero, P. Cassoni, M. Grangetto, Unitopatho, a labeled histopathological dataset for colorectal polyps classification and adenoma dysplasia grading, 2021, doi:10.48550/ARXIV.2101.09991.

[31] W. Hu, C. Li, X. Li, M.M. Rahaman, Y. Zhang, H. Chen, W. Liu, Y. Yao, H. Sun, N. Xu, X. Huang, M. Grzegorze, Ebhi:a new enteroscope biopsy histopathological h&e image dataset for image classification evaluation, 2022, doi:10.48550/ARXIV.2202.08552.

[32] K. Stacke, G. Eilertsen, J. Unger, C. Lundstrom, Measuring domain shift for deep learning in histopathology, IEEE J. Biomed. Health Inform. 25 (2) (2021) 325–336, doi:10.1109/JBHI.2020.3032060.

[33] H. Kang, D. Luo, W. Feng, S. Zeng, T. Quan, J. Hu, X. Liu, Stainnet: A Fast and robust stain normalization network, Front. Med. (Lausanne) 8 (2021) 2002, doi:10.3389/FMED.2021.746307/BIBTEX.

[34] M. Macenko, M. Niethammer, J.S. Marron, D. Borland, J.T. Woosley, X. Guan, C. Schmitt, N.E. Thomas, A method for normalizing histology slides for quantitative analysis, in: 2009 IEEE international symposium on biomedical imaging: from nano to macro, IEEE, 2009, pp. 1107–1110.

[35] E. Reinhard, M. Ashikhmin, B. Gooch, P. Shirley, Color transfer between images, IEEE Comput. Graph. Appl. 21 (5) (2001) 34–41, doi:10.1109/38.946629.

[36] M.T. Shaban, C. Baur, N. Navab, S. Albarqouni, Staingan: Stain style transfer for digital histological images, in: 2019 Ieee 16th international symposium on biomedical imaging (Isbi 2019), IEEE, 2019, pp. 953–956.

[37] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, K. Steiger, A.M. Schlitter, I. Esposito, N. Navab, Structure-Preserving color normalization and sparse stain separation for histological images, IEEE Trans. Med. Imag. 35 (8) (2016) 1962–1971, doi:10.1109/TMI.2016.2529665.

[38] A. Das, M.N. Mohanty, P.K. Mallick, P. Tiwari, K. Muhammad, H. Zhu, Breast cancer detection using an ensemble deep learning method, Biomed. Signal Process. Control 70 (August) (2021) 103009, doi:10.1016/j.bspc.2021.103009.

[39] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A ConvNet for the 2020s (2022). http://arxiv.org/abs/2201.03545.

[40] J. Yang, C. Li, J. Gao, Focal Modulation Networks (2022). http://arxiv.org/abs/2203.11926.

[41] K. He, X. Zhang, S. Ren, J. Sun, et al., Deep residual learning for image recognition, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2016-December (2015) 770–778, doi:10.48550/arxiv.1512.03385.

[42] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks. http://code.google.com/p/cuda-convnet/.

[43] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-Scale image recognition, 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings (2014), doi:10.48550/arxiv.1409.1556.

[44] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, et al., Going deeper with convolutions, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 07-12-June-2015 (2014) 1–9, doi:10.48550/arxiv.1409.4842.

[45] D.T. Nguyen, M.B. Lee, T.D. Pham, G. Batchuluun, M. Arsalan, K.R. Park, Enhanced image-Based endoscopic pathological site classification using an ensemble of deep learning models, Sensors (Basel) 20 (21) (2020) 1–24, doi:10.3390/S20215982.

[46] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, 2020, doi:10.48550/ARXIV.2010.11929.

[47] S. Yengec-Tasdemir, Implementation of Improved Classification of Colorectal Polyps on Histopathological Images with Ensemble Learning and

Stain Normalization, 2023, (https://github.com/senabyengec/Classification-of-Colorectal-Polyps-on-Histopathological-Images). [Online; accessed 06-Feb-2023].
[48] P. Byfield, StainTools, 2019, (https://github.com/Peter554/StainTools). Accessed: 01-05-2022.
[49] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: visual explanations from deep networks via gradient-based localization, Int. J. Comput. Vis. 128 (2) (2019) 336–359, doi:10.1007/s11263-019-01228-7.
[50] S.-j. Byeon, J. Park, Y.A. Cho, B.-J. Cho, Automated histological classification for digital pathology images of colonoscopy specimen via deep learning, Sci. Rep. 12 (1) (2022) 12804.
[51] S.H. Kassani, P.H. Kassani, M.J. Wesolowski, K.A. Schneider, R. Deters, Classification of histopathological biopsy images using ensemble of deep learning networks, in: CASCON 2019 Proceedings - Conference of the Centre for Advanced Studies on Collaborative Research - Proceedings of the 29th Annual International Conference on Computer Science and Software Engineering, 2020, pp. 92–99.
[52] J. Amann, A. Blasimme, E. Vayena, D. Frey, V.I. Madai, P. Consortium, Explainability for artificial intelligence in healthcare: a multidisciplinary perspective, BMC Med. Inform. Decis. Mak. 20 (2020) 1–9.
[53] A. Singh, S. Sengupta, V. Lakshminarayanan, Explainable deep learning models in medical image analysis, J. Imag. 6 (6) (2020) 52.
[54] S. Pereira, R. Meier, V. Alves, M. Reyes, C.A. Silva, Automatic brain tumor grading from MRI data using convolutional neural networks and quality assessment, Springer, 2018, pp. 106–114.
[55] R. Zhang, J. Zhu, S. Yang, S. Hosseini, A. Genovese, L. Chen, C. Rowsell, S. Damaskinos, S. Varma, K.N. Plataniotis, et al., HISTOKT: CROSS KNOWLEDGE TRANSFER IN COMPUTATIONAL PATHOLOGY, Technical Report, https://github.com/mahdihosseini/HistoKT.
[56] D. Perlo, E. Tartaglione, L. Bertero, P. Cassoni, M. Grangetto, et al., Dysplasia grading of colorectal polyps through convolutional neural network analysis of whole slide images, Technical Report