Generalizing Convolution Neural Networks on Stain Color Heterogeneous Data for Computational Pathology

Amjad Khan[a,b], Manfredo Atzori[b], Sebastian Otálora[b], Vincent Andrearczyk[b], and Henning Müller[b]

[a]Institute of Information Systems, University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland
[b]Institute of Pathology, University of Bern, CH-3008 Bern, Switzerland

ABSTRACT

Hematoxylin and Eosin (H&E) are one of the main tissue stains used in histopathology to discriminate between nuclei and extracellular material while performing a visual analysis of the tissue. However, histopathology slides are often characterized by stain color heterogeneity, due to different tissue preparation settings at different pathology institutes. Stain color heterogeneity poses challenges for machine learning-based computational analysis, increasing the difficulty of producing consistent diagnostic results and systems that generalize well. In other words, it is challenging for a deep learning architecture to generalize on stain color heterogeneous data, when the data are acquired at several centers, and particularly if test data are from a center not present in the training data. In this paper, several methods that deal with stain color heterogeneity are compared regarding their capability to solve center-dependent heterogeneity. Systematic and extensive experimentation is performed on a normal versus tumor tissue classification problem. Stain color normalization and augmentation procedures are used while training a convolutional neural networks (CNN) to generalize on unseen data from several centers. The performance is compared on an internal test set (test data from the same pathology institutes as the training set) and an external test set (test data from institutes not included in the training set). This also allows to measure generalization performance. An improved performance is observed when the predictions of the two best-performed stain color normalization methods with augmentation are aggregated. An average AUC and F1-score on external test are observed as 0.892 ± 0.021 and 0.817 ± 0.032 compared to the baseline 0.860 ± 0.027 and 0.772 ± 0.024 respectively.

Keywords: Histopathology, stain color heterogeneity, normalization, data augmentation, machine learning, computational pathology, CNN, generalization

1. INTRODUCTION

Histopathology is increasingly getting digital and with the digitization it is also moving towards computer-assisted solutions using machine learning on these large images. Visual inspection is currently the clinical standard in the field. However, computer-aided diagnosis tools can be particularly useful when visual inspection is time-consuming or when many quantitative parameters are requested to complete a diagnosis. For instance, analyzing tumor buddings as a prognostic marker in H&E stained images requires extensive visual studies.[1] The diagnostic outcome of a computer-aided system should not be related to color heterogeneity (such as differences due to center-dependent H&E staining procedures). However, image variability exists due to the variations in the thickness of the specimen, staining chemicals and properties of digital scanners (as shown in Fig. 1). These variations are then reflected in the diagnosis, and it is challenging to produce identical diagnostic results of the same tissue prepared with different pathology settings.[2]

Providing a correct diagnosis when dealing with stain color heterogeneous data is a challenging research topic. Various approaches have been developed to homogenize the stain color to generalize computational analysis.[3–8] Deep learning based solutions for various diagnostic tasks in histopathology have been extensively

Further author information: (Send correspondence to A.K.)
A.K.: E-mail: amjad.khan@pathology.unibe.ch

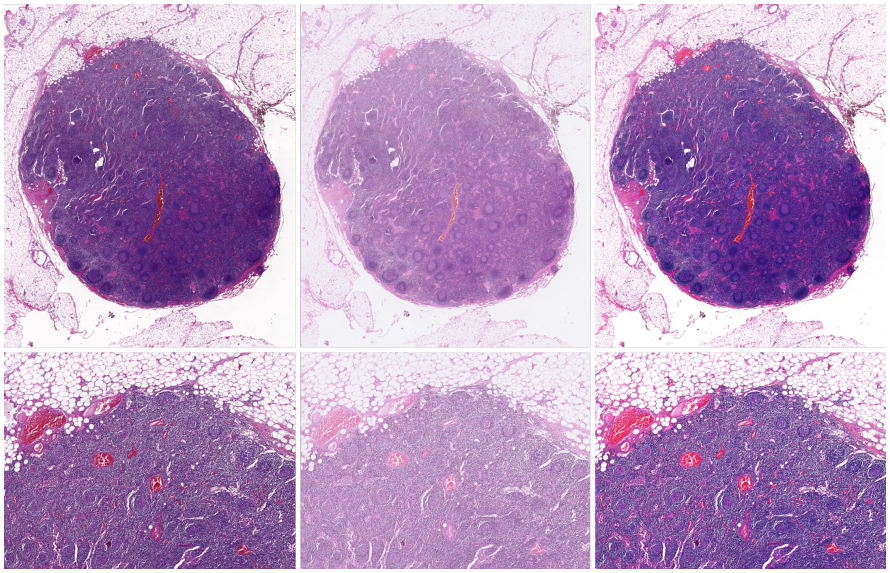

Figure 1: Visible stain color heterogeneity when the same tissue slide is scanned by three different scanners (top row) and zoom in version of the scans (bottom row).

explored.[9–12] Stain heterogeneity in digital pathology does not only affect the computational analysis performed by the research community but it is also a problem for the pathologists when they perform visual analysis on the whole slide images (for instance in telepathology). Such a problem is highlighted in,[13] where extensive experiments are performed to calibrate display systems using different color filters for H&E staining. In this study, staining, the thickness of the specimen, scanner and scanning process, viewing software, and displaying systems were considered responsible for color heterogeneity. The display systems study via the Macbeth color chart was helpful to calibrate the color of the displays across the entire department to make the visual effects homogeneous in histopathology data. Stain normalization in histopathology is not new to the digital image processing domain, especially when it comes to color variation in the image due to incandescent illumination, where it is essential to bring different images of the same scene to a standardized color distribution. This problem was already identified in,[4] where the authors suggested to transfer the color distribution of data to the color characteristics of one standard reference image among the data. The experimental results presented in[4] were mostly focused on outdoor scenes; however, due to the simplicity of the method, it has been used for histopathology slide normalization as well. In,[4] the authors presented a method based on the color distribution model in each channel of the *Lab* (L is lightness and a,b are color components) color space.

In digital pathology, the digitizing process of histopathlogy slides is quite similar to image acquisition in other domains. However, digitized slides contain variable microscopic information depending on the magnification used. In the slide scanning process, the appearance of the stain depends on the intensity absorbed by the tissue and it further depends on the amount of the stain added to the tissue and its handling and storage methods. The corresponding optical density describes the linear relationship between the stains and the absorbed intensity in the tissue for stain normalization using a deconvolution model.[3] In this method, the stain colors were estimated using a singular value decomposition (SVD) matrix. Then, source image values are mapped to the target matching through linear per channel normalization based on the 99th percentile. Furthermore, the method modifies the color distribution of both source and target images, which is undesirable in some cases where a suitable reference image is used to map the characteristics to the rest of the data. In order to overcome the problems faced by the method presented in,[3,4] a non-linear mapping of channel statistics based on the normalization method from source to target images is introduced in.[5] The method was mainly focused on the estimated stain matrix, color deconvolution, and reconstruction steps. The stain matrix estimation was performed by color classification. In the color classification, a Relevance Vector Machine (RVM) was trained on the red-green-blue color model. The significance of the method depends on the robust deconvolution matrix estimation and mapping function. The

color deconvolution separates the variation of each stain to correct it independently. However, the pre-trained RVM color classifier makes the method unstable in the test cases that deviate from the train cases due to varying dye color.

In,[6] the authors proposed a color standardizing technique for whole slide images by using the color and spatial information to classify the pixels into stain components. The density and chromatic distribution of the data in the hue-saturation-density (HSD) color space is aligned with a template slide. However, the performance of the method in the new data relies on expert opinion about the chosen reference template slide by considering the color and cellular information into account. In[14] a contrast limited adaptive histogram equalization(CLAHE) method is illustrated, based on intensity centering the model on bringing the color distribution to the center point within aggregated data. The method avoids the reference and target image statistics. However, the histogram equalization is limited to the spatial dependency of pixels. In order to preserve the biological structural information while performing the color normalization tasks[15] presented the structure-preserving color normalization. In this method, the stain density map was considered as sparse and non-negative. In sparsity, it was assumed that the biological material occupies exactly one given pixel location but not several. Similarly, the non-negativity describes that either a biological material is absorbing the light or not, and optical density cannot be negative. Based on the above assumptions, the color appearance and stained density matrices for both source and target images were estimated for color transformation.

Stain color normalization techniques can cope with the variability of the stain and the appearance of the digital histopathology slides for visual observations. The stain heterogeneity is also dealt with when usin machine learning-based approaches. These techniques focus on generalization improvement in the computational analysis by considering the features learned along with color information during the normalization process. In this context, deep convolution feature-aware normalization was presented in.[16] The study mainly focused on the visually relevant deep image features and style transfer. The feature-aware normalization was inspired by batch normalization[17] and long term memory[18] mechanisms. The method performed pixel-wise transformation based on features contained in the plasma or nucleus in the tissue. The color was treated as a form of a style to integrate into the network by shifting and scaling parameters of batch normalization layers. A pretrained VGG19 architecture mainly performed the features extraction process in both reference and source images. The mean and variance of the reference image features were used for color normalization by shifting and scaling network parameters. Similarly,[19] normalized the histopathology images by adopting the deep learning model that has been used on natural scenes colorization.[20] ResNet-v3 was used to extract the features from the images, and then these features were fused with the encoder-decoder model. The model was trained to estimate *ab* color values of *Lab* color space by minimizing the mean squared distance error between actual and estimated values.

In,[21] an unsupervised stain normalization method was introduced where the sparse auto-encoders were used to normalize moving images to a template image. The pixels were clustered by using k-means according to the respective tissue partitions in the sparse auto encoded feature space. Then, the color distribution of each partition in the moving data was aligned with its respective color distribution of the template. Finally, a histogram equalization was performed across the color channels. The method was used on several data sets of various tissue types with heterogeneity due to their domains or scanners. In digital histopathology, the color normalization techniques often require the reference image to transfer the color characteristics to the other images. However,[22] proposed to use end-to-end generative adversarial networks (GANs) to transfer the stain style by eliminating the requirement of an expert to choose a reference image. The experimentation was performed on the MITOS-ATYPIA dataset, which is acquired from the same tissue section with two different scanners.[23] The method mainly consisted of two pairs of generators and discriminator to map the stain style of the images belonging to one domain to the other. Similarly, GAN architectures were also employed by[24] to transfer the stain style from source to target images. The conditional GANs were trained to learn both color distribution and histopathological patterns present in the Camelyon16 data set.[25] The images from two different centers were normalized to gray-scale, then a style generator was used to color the gray-scale images again to a standard stain style. Inspired by style transfer and generative learning methods,[26] presented a stain normalization technique by preserving the structural information. The method avoided relying on a single reference image. The matching was performed on stain statistics over the entire domain of images. Instead of pixel-level matching, the feature representation of the images was used to normalize them. The proposed network was divided into stain transfer and task-specific parts to perform both stain normalization and classification or segmentation tasks simultaneously. The stain

transfer network learned the probability distribution of the images of one domain by minimizing the adversarial loss function to map the input image to a stain normalized image. In contrast, the task-specific model was used to maximizing the likelihood of the input image according to a given task. The proposed technique was evaluated on three data sets of mitosis, colon, and ovary,[23,26,27] that contained color variability. Similarly,[28,29] used an adversarial training framework to subtract the features that arise from the origin of the tissue from the features used to classify the images, showing significant improvements when combined with color augmentation techniques. Apart from stain transfer and feature-based stain normalization, the generalization in convolution neural networks for computational pathology can be improved by data augmentation methods. In,[30] a data augmentation based technique was developed to improve the generalization of the convolutional network for histopathology data. Each patch of the train set was modified in terms of hematoxylin, eosin, and residual channels, then a combination of rotation, color stain, scaling, elastic deformation, image enhancement, blurring, additive Gaussian noise was used as data augmentation techniques. Extensive experimentation was conducted to improve the generalization performance on different data sets of various tissues from multiple centers.[31–33] Finally, a recent review[34] is considered a reference for further analyses on various global, supervised, and unsupervised color normalization methods. This paper focuses on systematic and extensive experimentation of CNN generalization improvement when dealing with stain color heterogeneous data by minimizing the effects of stain color heterogeneity (mainly when the data are acquired at different centers). Improving the computer-aided diagnosis system capability to deal with staining color heterogeneity can foster the development of more reliable algorithms to improve the quality and reduce the workload of pathologist's clinical routine.

2. MATERIAL AND METHODS

In this paper, several stain color normalization methods and data augmentation techniques are evaluated using a convolution neural network classifier to improve the generalization, especially on the external data. The overall work-flow is shown in Fig. 2 and each of the blocks is described in the following subsections.

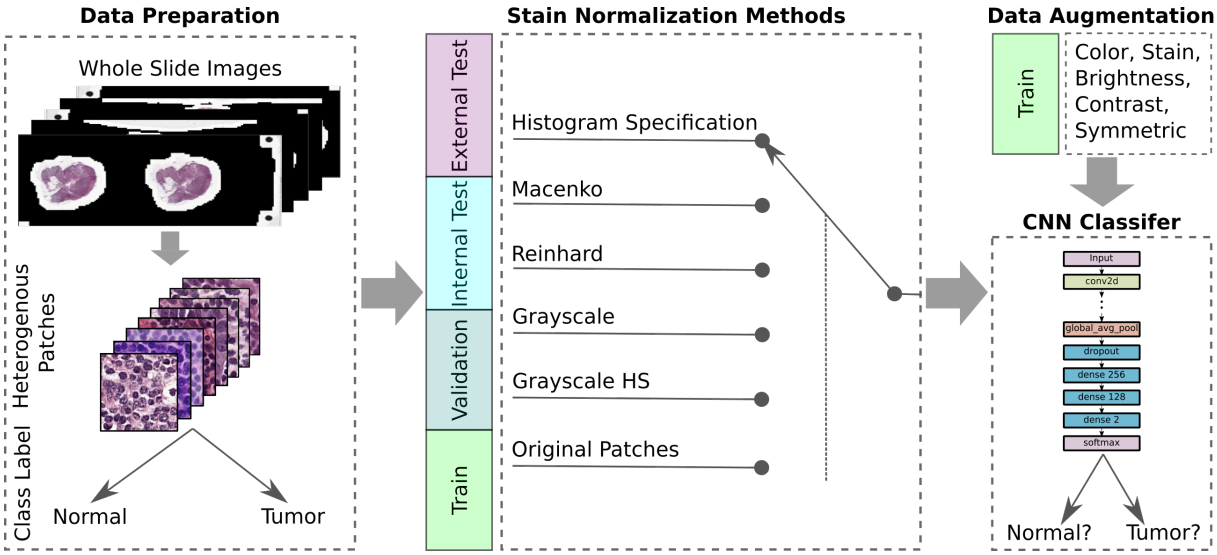

Figure 2: The block diagram presents the systematic experimentation approach to improve the generalization performance on stain heterogeneous data with main steps such as data preparation, stain normalization, data augmentation and classification of tissue regions with the CNN classifier.

2.1 Dataset

A dataset of 50 whole slide images (WSIs) of lymph node sections of the breast with local annotations of tumor regions from the CAMELYON 17 challenge is used.[25] WSIs were prepared at five pathology centers (10 slides from each center) and were scanned with three different scanners under pixel resolution of $0.23\mu m$ to $0.25\mu m$ and

provided in TIFF format. In order to train a CNN classifier, around 500 patches of 224×224 pixels are extracted from each tumor and normal region of every WSI (see Fig. 3). Then, the extracted patches from all slides are distributed into training, validation, internal test, and external test sets. In order to evaluate the generalization performance, this distribution is repeated each time by considering the patches from one pathology center as an external test set and removing them from all other sets. For extensive generalization evaluation of CNNs, five sub-data folds are obtained. A few examples of extracted patches from five pathology centers are presented in Fig. 4, in order to highlight stain color heterogeneity.

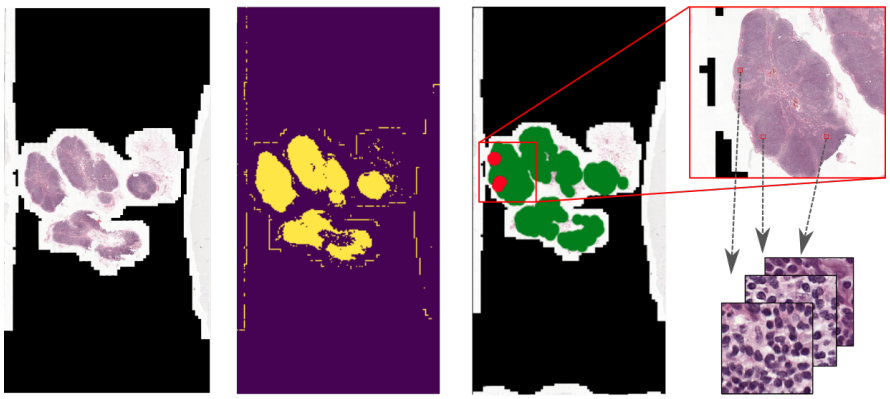

Figure 3: Patch extraction process, (a) Whole slide image, (b) segmented tissue mask, (c) annotated tumor lesions (red) and normal tissue (green) and (d) extracted patches.

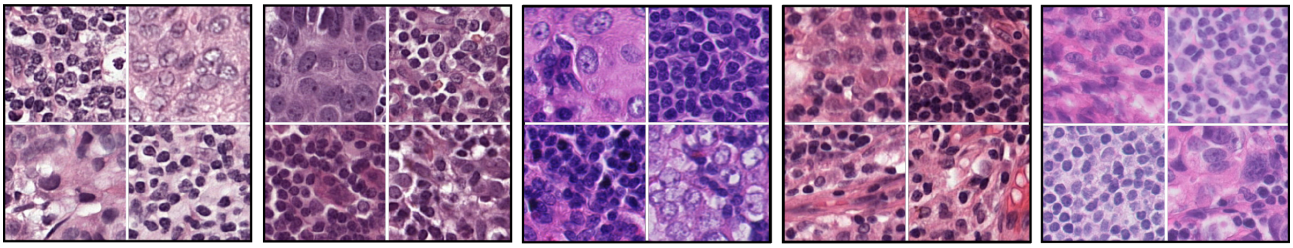

Figure 4: Stain color heterogeneity: tumor and normal tissue examples of the five CAMELYON17 pathology centers.

2.2 Stain Color Normalization

In order to minimize the stain color variations across data, various stain color normalization methods can be used to homogenize each fold (i.e., train, validation, internal test, and external test sets). The data partitions in each data fold are normalized to the same stain color distribution as of a template or target image. In this paper, three stain color normalization methods are evaluated on tumor versus normal tissue classification task.[35] Firstly, the histogram specification or matching is evaluated for stain color normalization in our data,[7,8] where histogram of each patch in the data is matched to the histogram of specified target or template image with the help of cumulative distribution function as given in Eq. 1 and Eq. 2.

$$cdf_{src(R,G,B)}(s_i) = \sum_{j=0}^{i} p_{src(R,G,B)}(s_j) \tag{1}$$

$$cdf_{tmp(R,G,B)}(t_i) = \sum_{j=0}^{i} p_{tmp(R,G,B)}(t_j) \tag{2}$$

i is total amount of gray level in each channel of the RGB (Red, Green, Blue) image, $cdf_{src(R,G,B)}(s_i)$ and $cdf_{tmp(R,G,B)}(t_i)$ are cumulative distribution functions of each gray level s_i and t_i in source and template image respectively. Similarly, $p_{src(R,G,B)}(s_j)$ and $p_{tmp(R,G,B)}(t_j)$ are representing the probability density function of each gray level s_i and t_i in source and template image respectively. The probability density functions are calculated from the histogram of both images, by considering the ratio of the frequency of the gray value to the total number of pixels in the channel. Finally, t_i in the template image is mapped to s_i in each source image for a uniform stain color distribution. Similarly, the stain color distribution among all the patches from train, validation, internal and external tests are specified to a single template image distribution in the RGB channels.

Second, the stain color normalization approach in [3] is evaluated where the normalization is performed in the H&E channels. All patches from each fold are mapped to a template image by estimating stain colors in optical density. Singular value decomposition (SVD) is used to get the optimal stain vectors from both input and template images to perform the linear per channel normalization based on the 99th percentile intensity values. Third, the stain color normalization of[4] is evaluated. The method is mainly based on a color distribution model in each channel of the *Lab* color space. Both source and template images are converted from RGB to *Lab*. Then mean and standard deviation of each channel of both images are calculated. The color distribution of the template image is transferred to the source images in the *Lab* color space by using the calculated mean and standard deviation as shown in Eq. 3.

$$I_{norm(L,a,b)} = \frac{\left[I_{src(L,a,b)} - \mu_{src(L,a,b)}\right] \times std_{tmp(L,a,b)}}{std_{src(L,a,b)} \times \mu_{tmp(L,a,b)}} \tag{3}$$

$I_{src(L,a,b)}$, $\mu_{src(L,a,b)}$ and $std_{src(L,a,b)}$ are the respective channels, mean and standard deviation values of source image in *Lab*. Similarly, $\mu_{tmp(L,a,b)}$ and $std_{tmp(L,a,b)}$ are representing mean and standard deviation of the corresponding template image respectively. Finally, the normalized image in *Lab* $I_{norm(L,a,b)}$ is converted back to RGB. Besides the above-mentioned stain color normalization methods, images are also converted to grayscale, and the grayscale histogram stretched versions are evaluated on the CNN model. Fig. 5 presents a few examples of the different patches from five centers with color variability, and then the above-mentioned normalization methods are used for uniform stain color distribution.

2.3 Data Augmentation

We hypothesized that a CNN can generalize to external data by learning more stain color variability. Therefore, the training samples are augmented in terms of color and stain, along with other symmetric data augmentation techniques. In order to augment the color variability, the training patches are modified in both the RGB and the HSV (Hue, Saturation, Value) color spaces by modifying, shuffling and shifting the channels. For RBG channels, the shifting values are between [-80:80, -45:45, -40:40] whereas the HSV channels are randomly shifted with a range of [-180:180, -20:20,-27:27]. The brightness and contrast variations are also produced with ranges between [-1.2:1.2] and [-0.9:0.9], respectively. Various H&E stain variations are produced by rescaling the stain vectors.[3] For the symmetric transformation, images are randomly rotated between [-100:100] degrees and flipped horizontally and vertically. A few examples of these augmentations are presented in Fig. 6 and Fig. 7.

2.4 Network and Training

In order to classify tumor versus normal tissue in the extracted patches, MobileNetV2 (showm in Fig. 8) is used.[36] The network is extended to two fully connected layers with 256 and 128 neurons. The probabilities of two classes are obtained with a layer of 2 neurons by using Softmax as an activation function. The network is pre-trained on ImageNet and then fine-tuned on our data by minimizing the cross-entropy loss with stochastic gradient descent as optimizer having an initial learning rate of 1×10^{-3} and dropping 50% neuron connections. The network is trained for 25 epochs with a learning rate halved every five epochs. A best-trained model is selected based on validation accuracy. This training procedure is repeated five times for each experiment, and average performance measures are reported.

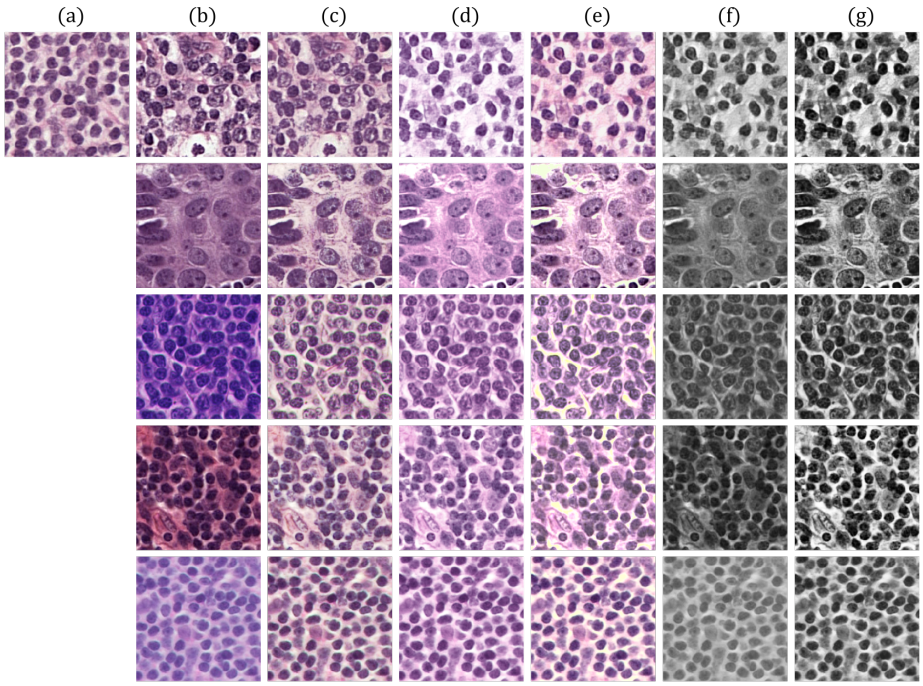

Figure 5: Stain color normalization, in the first column (a) A target or template image is used to distribute stain color homogeneously (b) original images with the help of different color normalization methods ((c) Histogram specification, (d) Macenko, (e) Reinhard) and also original images were homogenized to (f) gray-scale and (g) gray-scale histogram stretched images.

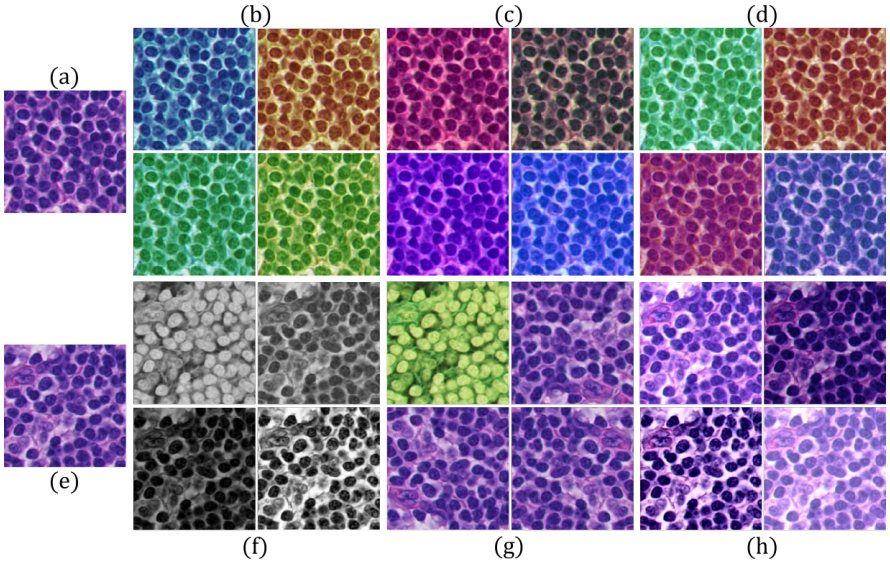

Figure 6: An example of data augmentation, on the original images (a) and (e), by (b) RGB channel shuffle, (c) RGB channel shifting, (d) HSV channel shifting, (f) brightness, contrast, inversion operations on gray version, (g) RGB inversion and symmetric operations and (h) brightness and contrast operations on RGB.

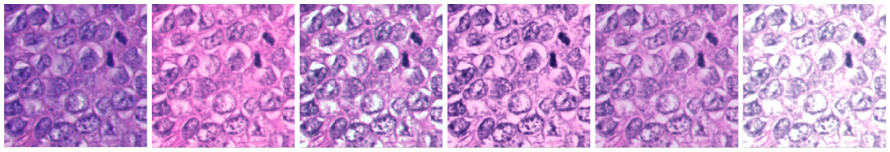

Figure 7: Stain augmentation, on (a) the original image by obtaining different stain augmented versions from (b) to (g).

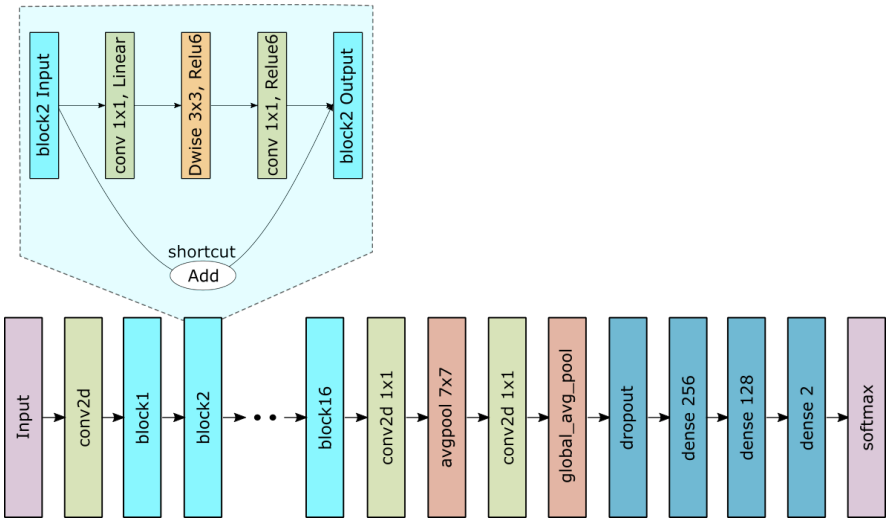

Figure 8: The architecture of aconvolutional neural network (MobileNetV2).

2.5 Evaluation and Statistical Metrics

In order to evaluate the generalization performance of the CNN on an external test set in each data fold, AUC (area under the receiver operating characteristics curve) and F1-score are used. Each normalization method with and without augmentation techniques is also evaluated by McNemar's statistical test applied to the class predictions.[37,38]The statistical comparison is performed based on p-value to obtain the most significant stain color normalization methods with or without augmentation on CNN classification when compare with the baseline (training of CNN classifier without any stain color normalization). This statistical analysis further helped in the ensemble process to choose the most significant normalization methods to fuse their probabilities for performance improvement.

3. RESULTS

An extensive and systematic experimentation approach is followed in each data fold by using normalization and augmentation techniques as well as their combination. A baseline is trained without any normalization method in each data fold. To compare the impact of having sample augmentations with each technique, the training samples are used with and without augmentation. The performance of each stain normalization method, along with the baseline, is presented in Table 1. The evaluation is based on averaged AUC and F1 score measures along with a standard deviation of 5 training repetitions with best-validated model weights for internal and external test sets of each data fold. In order to make the results more interpretive, the average performance scores across all five data folds for each of the techniques, including baseline are presented in Fig. 9(a)&(b). In the case of the experimentation without augmentation, the baseline obtained an average AUC across five data folds as 0.866 ± 0.026 and 0.833 ± 0.026 on internal and external test samples, respectively. The corresponding F1 score is measured as 0.780 ± 0.022 and 0.731 ± 0.051. Among the stain color normalization methods none scored higher than the baseline when evaluated on an internal test set with AUC. However, the histogram specification and

Reinhard outperformed on external test set with 0.864 ± 0.029, 0.856 ± 0.029 AUC and 0.777 ± 0.032, 0.765 ± 0.026 F1 scores respectively. It is evident from the results that all normalization techniques, including the baseline, show improved performance when evaluated on internal and external test samples with data augmentation. This suggests that including data-augmentations in training makes the training of the CNN models more robust to data heterogeneity. The average AUC across all data folds on the baseline is raised to 0.871 ± 0.027 and 0.860 ± 0.027 with corresponding F1 scores of 0.802 ± 0.025 and 0.772 ± 0.024 on internal and external test sets, respectively. Reinhard normalization with augmentation outperformed others, including the baseline with 0.878 ± 0.026, 0.877 ± 0.022 AUC, and 0.810 ± 0.023, 0.801 ± 0.024 F1 score in both internal and external test data respectively. After Reinhard, the histogram specification with data augmentation achieved 0.871 ± 0.022 AUC and a 0.782 ± 0.019 F1 score on the external test set. However, Macenko showed better performance than histogram specification on the internal test set, with 0.868 ± 0.026 AUC and 0.799 ± 0.023 F1 score.

McNemar's significance test[37,38] is performed to assess the most significant preprocessing settings on the classification compared to baseline on the obtained results, as shown in Table 1. The test is evaluated on the combinations of augmentation and normalization methods with their prediction on both internal and external test sets. The calculated p-values of each technique are presented in Table 1, where the last column contains the maximum p-value of internal and external test sets. From the statistical evaluation, histogram specification and Reinhard with and without augmentation on both test sets obtain an average p-value <0.017 at a significance level of 0.05. Smaller p-values then significance level showed that model on both methods predicted better than others. Therefore, predictions of both histogram specification and Reinhard are ensembled by fusing their probabilities through an element-wise multiplication, as shown inf Fig. 9(c).[39] Ensemble results on AUC and F1 measures are shown in Fig. 9(a)&(b). Average ensemble AUC scores without augmentation 0.881 ± 0.029, 0.882 ± 0.024 and with augmentation 0.885 ± 0.025, 0.893 ± 0.022 are obtained on internal and external tests respectively. The corresponding F1 scores are recorded as 0.822 ± 0.027, 0.795 ± 0.037 and 0.823 ± 0.024, 0.817 ± 0.032 respectively.

4. DISCUSSION

In this paper, the stain color heterogeneity in histopatholgy images was explored by minimizing its effects on a CNN-based classification task with the help of various stain color normalization techniques and data augmentation methods. For a combination of stain color normalization and data augmentation methods, a CNN classifier is trained to classify tumor and normal tissues acquired for the Camelyon17 challenge, containing heterogeneous stains from five histology centers. The prepared dataset is separated into five folds by considering each time the images from a center as an external source of patches for an external test evaluation. A baseline classifier is trained on each fold of data without any stain normalization or augmentation technique. As first experiments, patches from each fold are normalized in terms of stain color before passing to the CNN classifier to quantify their performance with the baseline. The stain color is normalized by histogram adaptations, namely Reinhard and Macenko. The experiments are also performed on grayscale and grayscale histogram stretched versions. By analyzing the obtained results on the normalization approaches, it is evident that the classifier performed well on external data when normalized with histogram specification and Reinhard. However, the Macenko and grayscale enhanced version shows almost identical performance on the external test set. In the second series of the experiments, the assessed data augmentation techniques are applied to the baseline and normalization based training. By introducing augmentation in training, we were able to improve the overall performance of the classifer. However, histogram specification and Reinhard again performed well on the external test set among other techniques. Then a statistical test is performed to obtain the best normalization methods along with augmentation to ensemble their probabilities. Where the McNemar's paired test also validated both histogram specification and Reinhard with a significance difference from the baseline at the class probability level. Therefore, in the third series of experiments, the probabilities of both outperformed methods along with augmentation are fused by element-wise multiplication. Interestingly, the CNN classifier learned different features on both normalization methods and results are improved on both internal and external test sets when their probabilities are fused. The best AUC and F1 score on external test are obtained as 0.893 ± 0.022 and 0.817 ± 0.032 respectively. One of the hypotheses was that the CNN classifier can learn better on a pattern instead of colors. In order to evaluate this hypothesis, the stain color was removed by converting the patches to grayscale versions. In a few cases the grayscale versions showed better performances. Overall results in the above

Table 1: Experimental results of tumor and normal tissue classification on the internal and external test sets with different normalization and augmentation techniques using a CNN. The values show the AUC and F1 scores averaged across 5 repetitions with standard deviation between the parenthesis along with p-values (paired McNemar's statistical test).

Data Fold	Normalization	Augmentation	Internal Test set		External Test set		p-value
			AUC	F1-score	AUC	F1-score	
1	Baseline	NA	0.847(0.016)	0.765(0.007)	0.860(0.006)	0.766(0.006)	-
	Histogram Specification	NA	**0.852(0.015)**	**0.773(0.018)**	0.862(0.014)	0.762(0.031)	<0.0001*
	Reinhard	NA	0.842(0.004)	0.751(0.016)	**0.871(0.007)**	**0.774(0.015)**	0.0206*
	Macenko	NA	0.841(0.005)	0.763(0.010)	0.827(0.021)	0.747(0.025)	0.0019*
	Grayscale	NA	0.834(0.020)	0.751(0.011)	0.845(0.011)	0.722(0.029)	0.0181*
	Grayscale-HS	NA	0.840(0.010)	0.773(0.012)	0.839(0.003)	0.765(0.003)	0.6870
	Baseline	CS,BCS	0.867(0.009)	0.809(0.005)	0.874(0.003)	0.782(0.009)	-
	Histogram Specification	CS,BCS	0.857(0.005)	0.791(0.010)	0.862(0.018)	0.780(0.018)	0.0016*
	Reinhard	CS,BCS	0.869(0.003)	0.798(0.008)	**0.889(0.003)**	**0.810(0.001)**	0.0080*
	Macenko	CS,BCS	**0.874(0.002)**	**0.819(0.006)**	0.854(0.003)	0.787(0.005)	0.0308*
	Grayscale	BCS	0.855(0.005)	0.748(0.050)	0.842(0.014)	0.710(0.073)	<0.0001*
	Grayscale-HS	BCS	0.850(0.006)	0.777(0.008)	0.844(0.003)	0.772(0.003)	0.1098
2	Baseline	NA	**0.876(0.010)**	0.780(0.017)	0.816(0.009)	0.734(0.010)	-
	Histogram Specification	NA	0.867(0.017)	**0.785(0.018)**	**0.838(0.011)**	**0.748(0.006)**	<0.0001*
	Reinhard	NA	0.861(0.021)	0.775(0.010)	0.823(0.015)	0.746(0.005)	<0.0001*
	Macenko	NA	0.848(0.003)	0.759(0.003)	0.812(0.003)	0.734(0.008)	0.6863
	Grayscale	NA	0.860(0.003)	0.761(0.023)	0.814(0.005)	0.726(0.014)	0.2538
	Grayscale-HS	NA	0.855(0.003)	0.743(0.0292)	0.829(0.008)	0.692(0.030)	<0.0001*
	Baseline	CS,BCS	0.876(0.005)	0.795(0.005)	0.834(0.006)	0.751(0.008)	-
	Histogram Specification	CS,BCS	0.866(0.004)	0.785(0.013)	**0.852(0.006)**	0.751(0.015)	0.1383
	Reinhard	CS,BCS	**0.888(0.006)**	**0.814(0.005)**	0.843(0.002)	**0.767(0.001)**	<0.0001*
	Macenko	CS,BCS	0.881(0.005)	0.781(0.021)	0.832(0.006)	0.727(0.006)	0.5349
	Grayscale	BCS	0.868(0.007)	0.776(0.006)	0.843(0.002)	0.761(0.009)	0.0236*
	Grayscale-HS	BCS	0.863(0.003)	0.772(0.014)	0.844(0.006)	0.755(0.016)	0.4895
3	Baseline	NA	**0.862(0.006)**	0.772(0.015)	0.797(0.013)	0.646(0.023)	-
	Histogram Specification	NA	0.850(0.002)	**0.805(0.003)**	0.858(0.002)	**0.773(0.008)**	<0.0001*
	Reinhard	NA	0.853(0.006)	0.777(0.012)	**0.860(0.005)**	0.745(0.031)	<0.0001*
	Macenko	NA	0.829(0.011)	0.748(0.025)	0.824(0.026)	0.735(0.010)	0.0048*
	Grayscale	NA	0.847(0.007)	0.760(0.021)	0.828(0.003)	0.731(0.017)	0.5986
	Grayscale-HS	NA	0.851(0.002)	0.784(0.012)	0.830(0.011)	0.733(0.013)	0.1465
	Baseline	CS,BCS	0.866(0.003)	0.789(0.013)	0.830(0.011)	0.742(0.003)	-
	Histogram Specification	CS,BCS	0.863(0.001)	0.804(0.008)	**0.877(0.006)**	**0.798(0.003)**	0.0090*
	Reinhard	CS,BCS	**0.869(0.002)**	**0.812(0.004)**	0.872(0.004)	0.796(0.001)	0.0001*
	Macenko	CS,BCS	0.850(0.002)	0.799(0.002)	0.857(0.004)	0.787(0.007)	0.0390*
	Grayscale	BCS	0.857(0.014)	0.780(0.024)	0.827(0.014)	0.733(0.008)	0.3717
	Grayscale-HS	BCS	0.851(0.001)	0.797(0.008)	0.838(0.009)	0.757(0.012)	0.6867
4	Baseline	NA	0.839(0.010)	0.766(0.009)	0.839(0.028)	0.773(0.017)	-
	Histogram Specification	NA	0.835(0.008)	0.751(0.024)	**0.913(0.013)**	**0.831(0.011)**	0.0665
	Reinhard	NA	**0.845(0.011)**	**0.786(0.011)**	0.883(0.013)	0.807(0.006)	<0.0001*
	Macenko	NA	0.810(0.006)	0.717(0.043)	0.867(0.043)	0.782(0.033)	0.0048*
	Grayscale	NA	0.815(0.018)	0.736(0.048)	0.876(0.007)	0.759(0.045)	0.5986
	Grayscale-HS	NA	0.833(0.012)	0.739(0.020)	0.898(0.010)	0.806(0.016)	0.1465
	Baseline	CS,BCS	0.835(0.003)	0.777(0.005)	0.857(0.005)	0.799(0.006)	-
	Histogram Specification	CS,BCS	0.829(0.015)	0.733(0.040)	0.890(0.024)	0.786(0.032)	<0.0001*
	Reinhard	CS,BCS	**0.848(0.001)**	**0.781(0.019)**	0.900(0.003)	0.832(0.003)	0.0154*
	Macenko	CS,BCS	0.834(0.001)	0.771(0.003)	0.870(0.001)	0.812(0.003)	0.1813
	Grayscale	BCS	0.827(0.009)	0.736(0.033)	0.897(0.016)	0.801(0.004)	<0.0001*
	Grayscale-HS	BCS	0.842(0.010)	0.781(0.009)	**0.902(0.034)**	**0.838(0.040)**	0.0998*
5	Baseline	NA	0.905(0.005)	0.819(0.002)	**0.852(0.014)**	0.736(0.052)	-
	Histogram Specification	NA	0.903(0.003)	0.828(0.007)	0.847(0.008)	**0.766(0.003)**	0.0098*
	Reinhard	NA	**0.920(0.006)**	**0.849(0.006)**	0.841(0.011)	0.752(0.005)	0.0341*
	Macenko	NA	0.885(0.014)	0.802(0.006)	0.830(0.003)	0.719(0.034)	0.0122*
	Grayscale	NA	0.887(0.004)	0.808(0.001)	0.728(0.043)	0.550(0.083)	0.1282
	Grayscale-HS	NA	0.897(0.005)	0.824(0.009)	0.811(0.009)	0.705(0.035)	0.2006
	Baseline	CS,BCS	0.911(0.004)	0.841(0.007)	0.867(0.008)	0.786(0.005)	-
	Histogram Specification	CS,BCS	0.901(0.002)	0.830(0.003)	0.872(0.021)	0.794(0.021)	<0.0001*
	Reinhard	CS,BCS	**0.918(0.007)**	**0.844(0.005)**	**0.882(0.002)**	**0.801(0.007)**	0.0660*
	Macenko	CS,BCS	0.900(0.003)	0.823(0.003)	0.851(0.014)	0.775(0.011)	0.0291*
	Grayscale	BCS	0.892(0.004)	0.822(0.002)	0.804(0.005)	0.719(0.009)	<0.0001*
	Grayscale-HS	BCS	0.898(0.001)	0.838(0.003)	0.857(0.013)	0.766(0.013)	0.2979

Note: NA: No Augmentation, CS: Color and Stain, BCS: Brightness, Contrast and Symmetric, HS: Histogram Stretched
*p-value<0.05 (paired McNemar's Test)

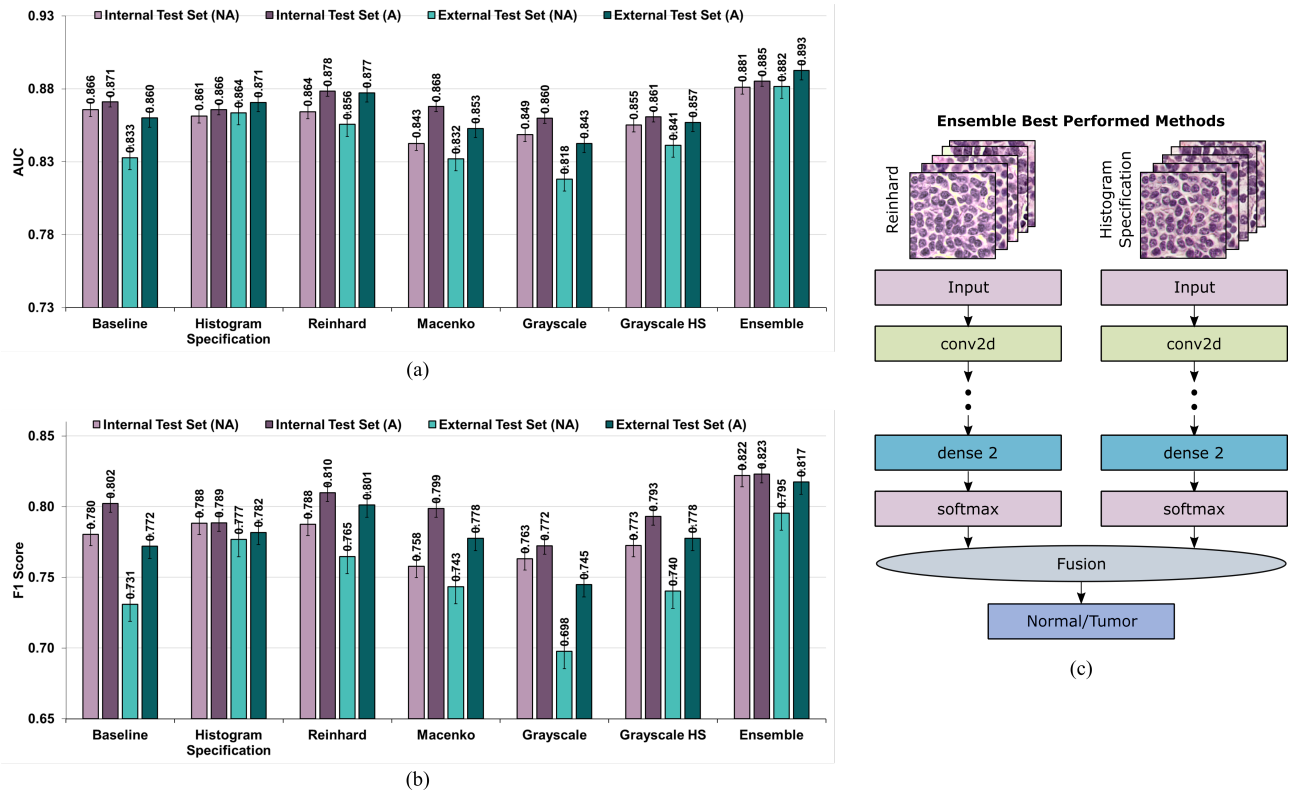

Figure 9: Average (a) AUC and (b) F1-score (and standard deviations) across 5 sub-data folds of each stain color normalization method with augmentation and without augmentation comparing with the baseline individually as well as the results when (c) ensemble technique is used to fuse two best-performed methods at prediction level. (Note: A:Augmentation, NA:No Augmentation, HS:Histogram Stretched.)

experiments make evident that the stain colors are important and effect the CNN classification based decisions along with morphological patterns.

5. CONCLUSIONS AND FUTURE WORK

In this paper, extensive and systematic experimentation is presented to improve the CNN generalization on heterogeneous H&E stained pathology images for breast tumor versus normal tissue classification tasks. Various stain color normalization methods, augmentation techniques and combinations of both are investigated while training the CNN on a stain color heterogeneous dataset. The performance is compared on the test data from the same pathology institutes of the training set (internal test set) and test data from institutes not included in the training set (external test set). Experimental results show better performance when stain color normalization and augmentation are used together on external test samples. However, results further improved when the two best-performing methods are fused. In future, this work can be extended further to validate the best-performing methods on other types of tissue with H&E stain color heterogeneity such as colorectal and prostate with the aim of cancer classification or grading. In present work, normalization is limited to a single template image. However, in future work several template images can be used to analyze the robustness of the normalization method as well as by including more normalization techniques to the comparison.

6. ACKNOWLEDGMENTS

This project received funding from the European Unions Horizon 2020 research and innovation program under grant agreement No. 825292. Amjad Khan holds an Erasmus Mundus master's scholarship. This work is part of

his master thesis during Erasmus Mundus Joint Master Degree in Medical Imaging and Applications. The authors are grateful for the management and the professors of all three universities in the consortium (the University of Burgundy-France, University of Cassino-Italy and University of Girona-Spain) for their support during the master program. The authors are also thankful to the University of Applied Sciences Western Switzerland (HES-SO) for facilitating the work during the thesis period. Special thanks to all the colleagues at HES-SO for very useful discussions during this work.

REFERENCES

[1] Lugli, A., Kirsch, R., Ajioka, Y., Bosman, F., Cathomas, G., Dawson, H., El Zimaity, H., Fléjou, J. F., Hansen, T. P., Hartmann, A., Kakar, S., Langner, C., Nagtegaal, I., Puppa, G., Riddell, R., Ristimäki, A., Sheahan, K., Smyrk, T., Sugihara, K., Terris, B., Ueno, H., Vieth, M., Zlobec, I., and Quirke, P., "Recommendations for reporting tumor budding in colorectal cancer based on the International Tumor Budding Consensus Conference (ITBCC) 2016," *Modern Pathology* **30**, 1299–1311 (sep 2017).

[2] Janowczyk, A. and Madabhushi, A., "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases," *Journal of pathology informatics* **7** (2016).

[3] Macenko, M., Niethammer, M., Marron, J. S., Borland, D., Woosley, J. T., Guan, X., Schmitt, C., and Thomas, N. E., "A method for normalizing histology slides for quantitative analysis," in [*Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on*], 1107–1110, IEEE (2009).

[4] Reinhard, E., Ashikhmin, M., Gooch, B., and Shirley, P., "Color transfer between images," *IEEE Computer Graphics and Applications* **21**(5), 34–41 (2001).

[5] Khan, A. M., Rajpoot, N., Treanor, D., and Magee, D., "A Nonlinear Mapping Approach to Stain Normalization in Digital Histopathology Images Using Image-Specific Color Deconvolution," *IEEE Transactions on Biomedical Engineering* **61**, 1729–1738 (jun 2014).

[6] Ehteshami Bejnordi, B., Litjens, G., Timofeeva, N., Otte-Holler, I., Homeyer, A., Karssemeijer, N., and van der Laak, J. A., "Stain Specific Standardization of Whole-Slide Histopathological Images," *IEEE Transactions on Medical Imaging* **35**, 404–415 (feb 2016).

[7] Gonzalez, R. C. and Woods, R. E., [*Digital Image Processing (3rd Edition)*], Prentice-Hall, Inc., Upper Saddle River, NJ, USA (2006).

[8] Coltuc, D., Bolon, P., and Chassery, J.-M., "Exact histogram specification," *IEEE Transactions on Image Processing* **15**, 1143–1152 (may 2006).

[9] Xu, Y., Jia, Z., Wang, L.-B., Ai, Y., Zhang, F., Lai, M., and Chang, E. I.-C., "Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features," *BMC Bioinformatics* **18**, 281 (dec 2017).

[10] Li, Z., Hu, Z., Xu, J., Tan, T., Chen, H., Duan, Z., Liu, P., Tang, J., Cai, G., Ouyang, Q., Tang, Y., Litjens, G., and Li, Q., "Computer-aided diagnosis of lung carcinoma using deep learning - a pilot study," **1** (2018).

[11] Otálora, S., del Toro, O. J., Atzori, M., Khan, A., Andrearczyk, V., and Müller, H., "A systematic comparison of deep learning strategies for weakly supervised gleason grading," in [*Medical Imaging 2020: Digital Pathology.To appear*], International Society for Optics and Photonics (2020).

[12] del Toro, O. J., Atzori, M., Otlora, S., Andersson, M., Eurn, K., Hedlund, M., Rnnquist, P., and Mller, H., "Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade Gleason score," in [*Medical Imaging 2017: Digital Pathology*], Gurcan, M. N. and Tomaszewski, J. E., eds., **10140**, 165 – 173, International Society for Optics and Photonics, SPIE (2017).

[13] Yagi, Y., "Color standardization and optimization in Whole Slide Imaging," *Diagnostic Pathology* **6**, S15 (dec 2011).

[14] Tam, A., Barker, J., and Rubin, D., "A method for normalizing pathology images to improve feature extraction for quantitative pathology," *Medical Physics* **43**, 528–537 (jan 2016).

[15] Vahadane, A., Peng, T., Sethi, A., Albarqouni, S., Wang, L., Baust, M., Steiger, K., Schlitter, A. M., Esposito, I., and Navab, N., "Structure-Preserving Color Normalization and Sparse Stain Separation for Histological Images," *IEEE Transactions on Medical Imaging* **35**, 1962–1971 (aug 2016).

[16] Bug, D., Schneider, S., Grote, A., Oswald, E., Feuerhake, F., Schüler, J., and Merhof, D., [*Context-Based Normalization of Histological Stains Using Deep Convolutional Features*], Springer International Publishing, Cham (2017).

[17] Ioffe, S. and Szegedy, C., "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in [*Proceedings of the 32nd International Conference on International Conference on Machine Learning*], 448—-456, JMLR.org (feb 2015).

[18] Hochreiter, S. and Schmidhuber, J., "Long Short-Term Memory," *Neural Computation* **9**, 1735–1780 (nov 1997).

[19] Samsi, S., Jones, M., Kepner, J., and Reuther, A., "Colorization of H&E stained tissue using Deep Learning," in [*2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*], 640–643, IEEE (jul 2018).

[20] Baldassarre, F., González Morín, D., and Rodés-Guirao, L., "Deep Koalarization: Image Colorization using CNNs and Inception-Resnet-v2," (2017).

[21] Janowczyk, A., Basavanhally, A., and Madabhushi, A., "Stain normalization using sparse autoencoders (stanosa): Application to digital pathology," *Computerized Medical Imaging and Graphics* **57**, 50–61 (2017).

[22] Shaban, M. T., Baur, C., Navab, N., and Albarqouni, S., "StainGAN: Stain Style Transfer for Digital Histological Images," **1**, 1–8 (2018).

[23] Roux, L., Racoceanu, D., Capron, F., Calvo, J., Attieh, E., Naour, G. L., and Gloaguen, A., "Mitos & atypia," (2014).

[24] Cho, H., Lim, S., Choi, G., and Min, H., "Neural Stain-Style Transfer Learning using GAN for Histopathological Images," **80**, 1–10 (oct 2017).

[25] Litjens, G., Bandi, P., Ehteshami Bejnordi, B., Geessink, O., Balkenhol, M., Bult, P., Halilovic, A., Hermsen, M., van de Loo, R., Vogels, R., Manson, Q. F., Stathonikos, N., Baidoshvili, A., van Diest, P., Wauters, C., van Dijk, M., and van der Laak, J., "1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset," *GigaScience* **7** (jun 2018).

[26] Bentaieb, A. and Hamarneh, G., "Adversarial stain transfer for histopathology image analysis," *IEEE transactions on medical imaging* **37**(3), 792–802 (2018).

[27] Sirinukunwattana, K., Pluim, J. P. W., Chen, H., Qi, X., Heng, P.-A., Guo, Y. B., Wang, L. Y., Matuszewski, B. J., Bruni, E., Sanchez, U., Böhm, A., Ronneberger, O., Cheikh, B. B., Racoceanu, D., Kainz, P., Pfeiffer, M., Urschler, M., Snead, D. R. J., and Rajpoot, N. M., "Gland Segmentation in Colon Histology Images: The GlaS Challenge Contest," **1** (2016).

[28] Otálora, S., Atzori, M., Andrearczyk, V., Khan, A., and Müller, H., "Staining invariant features for improving generalization of deep convolutional neural networks in computational pathology," *Frontiers in Bioengineering and Biotechnology* **7**, 198 (2019).

[29] Lafarge, M. W., Pluim, J. P. W., Eppenhof, K. A. J., and Veta, M., "Learning domain-invariant representations of histological images," *Frontiers in Medicine* **6**, 162 (2019).

[30] Tellez, D., Balkenhol, M., Karssemeijer, N., Litjens, G., van der Laak, J., and Ciompi, F., "H and e stain augmentation improves generalization of convolutional networks for histopathological mitosis detection," in [*Medical Imaging 2018: Digital Pathology*], **10581**, 105810Z, International Society for Optics and Photonics (2018).

[31] Veta, M., Heng, Y. J., Stathonikos, N., Bejnordi, B. E., Beca, F., Wollmann, T., Rohr, K., Shah, M. A., Wang, D., Rousson, M., Hedlund, M., Tellez, D., Ciompi, F., Zerhouni, E., Lanyi, D., Viana, M., Kovalev, V., Liauchuk, V., Phoulady, H. A., Qaiser, T., Graham, S., Rajpoot, N., Sjöblom, E., Molin, J., Paeng, K., Hwang, S., Park, S., Jia, Z., Chang, E. I.-C., Xu, Y., Beck, A. H., van Diest, P. J., and Pluim, J. P., "Predicting breast tumor proliferation from whole-slide images: The TUPAC16 challenge," *Medical Image Analysis* **54**, 111–121 (jul 2019).

[32] Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Ehteshami Bejnordi, B., Lee, B., Paeng, K., Zhong, A., Li, Q., Zanjani, F. G., Zinger, S., Fukuta, K., Komura, D., Ovtcharov, V., Cheng, S., Zeng, S., Thagaard, J., Dahl, A. B., Lin, H., Chen, H., Jacobsson, L., Hedlund, M., Cetin, M., Halici, E., Jackson, H., Chen, R., Both, F., Franke, J., Kusters-Vandevelde, H., Vreuls, W., Bult, P., van Ginneken, B., van der Laak, J., and Litjens, G., "From Detection of Individual Metastases to Classification

of Lymph Node Status at the Patient Level: The CAMELYON17 Challenge," *IEEE Transactions on Medical Imaging* **38**, 550–560 (feb 2019).

[33] Kather, J. N., Weis, C.-A., Bianconi, F., Melchers, S. M., Schad, L. R., Gaiser, T., Marx, A., and Zöllner, F. G., "Multi-class texture analysis in colorectal cancer histology," *Scientific Reports* **6**, 27988 (sep 2016).

[34] Roy, S., kumar Jain, A., Lal, S., and Kini, J., "A study about color normalization methods for histopathology images," *Micron* **114**(July), 42–61 (2018).

[35] Byfield, P., "Staintools." https://github.com/Peter554/StainTools (2019).

[36] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C., "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in [*2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 4510–4520, IEEE (jun 2018).

[37] McNemar, Q., "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika* **12**, 153–157 (Jun 1947).

[38] Raschka, S., "Mlxtend: Providing machine learning and data science utilities and extensions to pythons scientific computing stack," *The Journal of Open Source Software* **3** (Apr. 2018).

[39] Liu, Y., Zhang, C., Cheng, J., Chen, X., and Wang, Z. J., "A multi-scale data fusion framework for bone age assessment with convolutional neural networks," *Computers in Biology and Medicine* **108**, 161–173 (may 2019).